DECISION SCIENCES INSTITUTE

An Empirical Investigation of the Baldrige Framework Using Applicant Scoring Data

Feng Mai
Stevens Institute of Technology
fmai@stevens.edu

Matthew W. Ford
Northern Kentucky University
fordmw@nku.edu

James R. Evans
University of Cincinnati
evansjr@ucmail.uc.edu

ABSTRACT

This paper seeks to overcome evaluative limitations of previous studies to provide a more decisive test of the causal relationships implied in the Baldrige Criteria for Performance Excellence (CPE). Path analysis of blinded applicant scoring data from 500 Baldrige Award applicants provided strong support for the CPE framework in its entirety, with all hypothesized relationships between constructs confirmed. However, analysis of sector-specific subsets of the data did not confirm all relationships, suggesting the possibility of industry-dependent performance excellence frameworks and raising new research questions to be explored.

KEYWORDS: Quality management and systems, Process improvement, Performance measurement, Path analysis, Structural equations modeling

INTRODUCTION

Since its inception in 1988, the Malcolm Baldrige National Quality Award Criteria for Performance Excellence (CPE) have evolved from a narrow focus on quality management practices to a broad framework for organizational performance excellence (Evans et al., 2012). Its broad scope and acceptance among manufacturing, service, and non-profit organizations suggests that the CPE framework is an effective standard for promoting performance excellence among organizations operating in various internal and external contexts (Bemowski & Stratton, 1995; Ettorre, 1996; Flynn & Saladin, 2006; Steeples, 1992). As the CPE have increased in prominence, scholars have sought to empirically validate the underlying conceptual models.

Revisions to the CPE during the course of the criteria's evolution have resulted in three generations of conceptually distinct frameworks (Flynn & Saladin, 2001): a first generation framework in place during the 1988 to 1991 Baldrige Award cycles, a second generation framework in place from 1992 to 1996, and a third generation framework introduced in 1997, which is still in place. Several studies aimed at validating the CPE framework were based on earlier generation frameworks (e.g., Meyer & Collier, 2001; Pannirselvam & Ferguson, 2001; Wilson & Collier, 2000), thereby limiting the relevance of a significant portion of research findings to the current CPE structure. Moreover, studies that have investigated the validity of the third generation framework have been unable to statistically confirm all hypothesized links between CPE constructs on a consistent basis, with non-significant paths sometimes amounting to 20% or more of total paths tested (e.g., Flynn & Saladin, 2001; Lee et al., 2003). While a review of the research by Bou-Llusar et al. (2009) found that empirical studies generally support the conceptual framework of the CPE, the general support that flows from extant empirical work has not translated into confirmation of the CPE model in its entirety. Thus, although general support for the framework's validity is evident, specific issues remain that merit additional research.

Confirmation of relationships between constructs of the CPE framework has been hindered in part by the latitude that researchers have taken in interpreting the conceptual framework of the CPE. This has resulted in a plethora of model configurations and a variety of constructs and relationships proposed for testing (e.g., Curkovic et al., 2000; Flynn & Saladin, 2001; He et al., 2011; Jayamaha et al., 2008; Karimi. et al., 2014; Lee et al., 2003). Studies attempting to replicate or extend the validity of previously tested models have been rare. Methodological limitations are also evident. Data have commonly been collected using survey questionnaires that have varied considerably in design (e.g., Curkovic et al., 2000; Ghosh et al., 2003; Goldstein & Schweikhart, 2002; Lee et al., 2003; Wilson & Collier, 2000). Some instruments have contained items specifically developed to reflect the CPE framework while others include questions sourced from other studies. Sample domains have usually been limited to industry-specific sectors such as manufacturing, healthcare, or education, thereby constraining assessment to narrow sets of evaluative conditions. Respondents have been sourced from various organizational groups ranging from senior executives to quality managers to line personnel which, given well-known attribution biases and errors associated with group sampling (Dearborn & Simon, 1958; Hambrick, 1981; Nisbett & Wilson, 1977) casts some doubt on the capability of respondents from different departments or hierarchical organizational levels to accurately assess the state of far reaching, complex organizational systems such as those represented in the CPE.

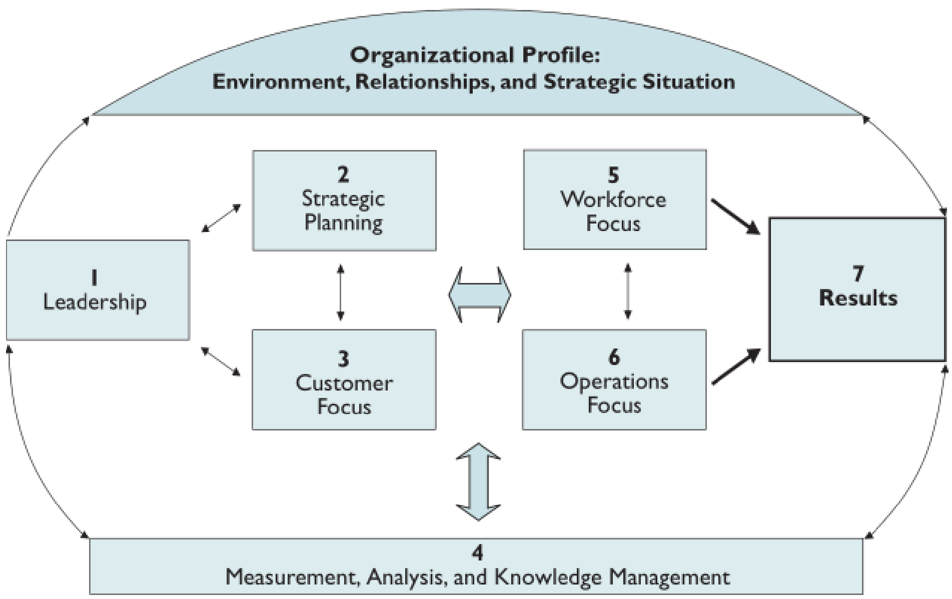
Because the CPE is employed both by managers as a means for self-assessment and performance improvement (Ford & Evans, 2001) and by researchers as an operationalization of theoretical total quality management (TQM) concepts (Bou-Llusar et al., 2009), it is important that the validity of all constructs and linkages proposed by the CPE framework be thoroughly examined. The limitations associated with past studies suggest that further that work is necessary in order to more fully understand and validate the framework.

This paper seeks to address past limitations to provide a more decisive empirical test of the CPE framework. We employ Flynn and Saladin's (2001) representation of the current (third generation) CPE framework for our test. In addition to offering a carefully developed representation of the CPE framework as a basis for evaluation, use of Flynn and Saladin's (2001) model enriches opportunity for cross study comparison that has eluded previous investigations. We test the model by analyzing primary scoring data from 500 organizations that applied for the Malcolm Baldrige National Quality Award since 1999 that were compiled and released by the Baldrige Program. Because these data span multiple industrial sectors and reflect ratings of expert observers with extensive skill in assessing organizational activities in light of the CPE framework, use of Baldrige applicant scoring data overcomes many methodological limitations of past studies and allows us to investigate the universal applicability of the framework across industries. The primary contribution of this research is an improved analytical backdrop that provides more favorable conditions for thorough empirical testing of the CPE framework.

LITERATURE REVIEW

As noted above, Flynn and Saladin (2001) separated the conceptual evolution of the CPE framework into three generations. Although its content and terminology have undergone minor revision since then, the third generation framework enacted in 1997 remains in effect today. As shown in Figure 1 (based on the 2013-2014 Criteria), the framework proposes six categories of interlinked organizational processes related to leadership, strategy development and implementation, customer focus, information and analysis, workforce focus, and operations focus (process management). Each category is subdivided into several items that focus on factors that organizations should address to function effectively in competitive environments. The six process categories are posited to influence a seventh multidimensional results category. We note that the current 2015-2016 version of the framework has not changed substantially, except for minor category title wording changes such as “Strategy,” “Customers”, “Workforce,” and “Operations.” Detailed scholarly reviews of the conceptual CPE framework have appeared on numerous occasions; readers are referred to Evans and Mai (2014) for a recent review.

Figure 1 Third Generation Baldrige framework



Several empirical studies have investigated the validity of the third generation CPE framework indirectly by examining novel combinations of the categorical constructs (Table 1). For example, Curkovic et al., (2002) and Jayamaha et al., (2011) reconfigured the seven categorical constructs into several second order factor models to evaluate the extent to which the CPE framework reflected theoretical TQM and business excellence constructs. Goldstein and Schweikhart (2002) assessed the predictive validity of the six CPE process categories relative to various outcome dimensions specified by the seventh results category. Although findings from these studies favor the validity of the CPE framework, support for the particular configuration of the seven CPE constructs as proposed by the CPE developers in Figure 1 is limited due to the scopes of the investigations. Confirming the complete arrangement of proposed constructs and relationships, sometimes called a “nomological network,” is critical to assessing total validity of a theoretical framework (Cronbach & Meehl, 1955).

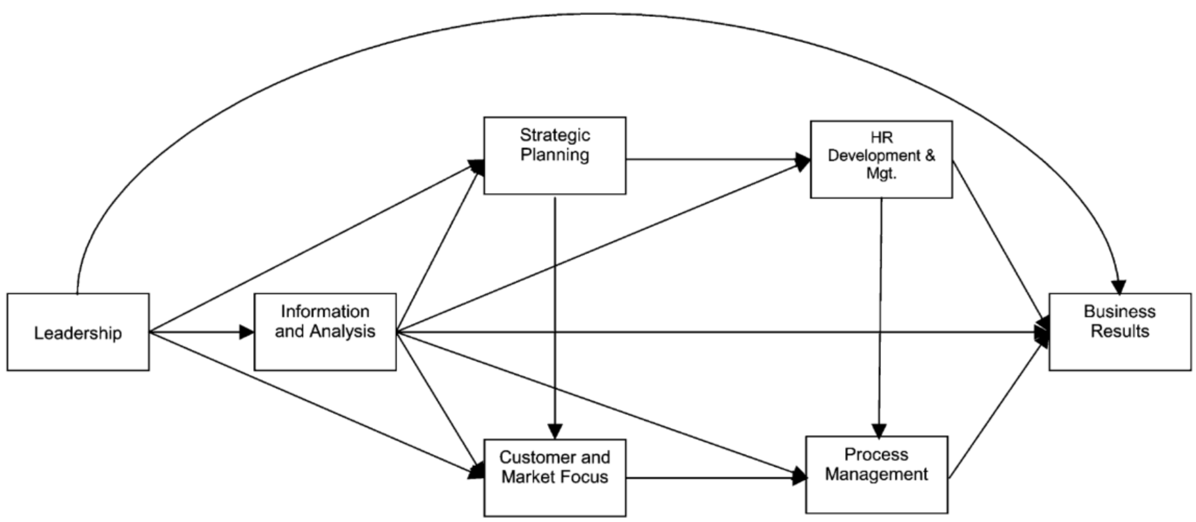
The remaining studies in Table 1 constitute direct examinations of the causal network of constructs and relationships expressed by the third generation CPE framework in. In each of these studies, the conceptual framework in Figure 1 was converted into a measurement model consisting of the seven interlinked CPE categorical constructs arranged in causal order. Model fit and significance were then usually tested using path analytical techniques and empirical data samples. However, model specification has varied between studies with no two investigations testing identical representations of the CPE framework pictured in Figure 1. Causal order of the CPE constructs has also differed between studies. For example, while the leadership construct has been positioned as the sole exogenous "driver" construct in several measurement models (e.g., Flynn & Saladin, 2001; Jayamaha et al., 2008; He et al., 2011), leadership has shared this position with the information and analysis construct in other studies (e.g., Lee et al., 2003).

Table 1: Empirical Studies Employing Third Generation CPE Framework

Study	CPE Basis	Sample	Analytical Approach	Linkages Confirmed/Proposed
Curkovic et al., (2000)	1997	Single respondent survey of 269 plant managers in US automotive industry	CPE in second order factor structural equation modeling (SEM) analysis of TQM	N.A.
Flynn & Saladin (2001)	1997	Multi-respondent survey of 164 manufacturing plants in Germany, Italy, Spain, UK, US	Path analysis of CPE causal model	12/15
Goldstein & Schweikhart (2002)	1999	Single respondent survey of quality managers at 220 US hospitals	Regression analysis of CPE process categories and results category	6/6
Ghosh et al. (2003)	2000	Single respondent survey of quality managers in 313 US manufacturing firms.	SEM analysis of CPE causal model	5/10
Lee et al., (2003)	2001	Single respondent survey of 109 Korean mid-upper level mfg mgrs	SEM analysis of CPE causal model	8/12
Badri et al., (2006)	2004	Multi-respondent survey of 224 administrators from 15 UAE colleges and universities	SEM analysis of CPE causal model	21/21
Jayamaha et al., (2008)	2005	Single respondent survey of 91 New Zealand organizations	SEM (PLS) analysis of CPE causal model	11/13
He et al., (2011)	2006	Single respondent survey from 2302 Chinese manufacturers and service firms	SEM analysis of CPE causal model	19/19
Jayamaha et al., (2011)	2009	Single respondent survey of 118 New Zealand organizations	CPE in second order factor SEM (PLS) model of business excellence	N.A.
Karimi et al., (2014)	2006	Applicant scoring data from 277 Baldrige Award applicants 2003-2006	Canonical correlation analysis of Baldrige drivers, systems, results	N.A.

The two-headed arrows shown to connect various CPE categories and groups of constructs in Figure 1 have been a source of confusion for researchers. Investigators have generally assumed the two-headed arrows to mean that the CPE developers were unsure of the direction of true causation between the constructs, leading the developers to default to the premise that all CPE categories were related (e.g., Meyer & Collier, 2001; Wilson & Collier, 2000). However, two-headed arrows and the non-causality that they imply create significant problems for path analytical methodology, and have lead researchers to specify nomological networks with directional causal order. When modeling the CPE framework, researchers have argued that the long articulated CPE proposition that “leadership drives the system that leads to results” justifies arranging the model in recursive, unidirectional order that proceeds through some configuration of the six organizational process constructs and ends with the seventh results construct. These model configurations contain no reciprocal causation or feedback loops. All path coefficients between constructs are hypothesized as positive. As demonstrated by Flynn and Saladin’s (2001) model (Figure 2), researchers have redrawn the third generation CPE to conform to the conventions of path analysis. All arrows are unidirectional and point toward the results end of the model. (The 2015-2016 version of the framework replaced the multiple arrows in the figure with a central theme of “integration” that implies that all the elements of the system are interrelated, leaving the causal paths open to interpretation.) The Baldrige categories have undergone some title revisions since 1997 (on which Figure 2 is based); thus, in our investigation, we will use the category names from Figure 1; that is, Leadership; Measurement, Analysis, and Knowledge Management; Strategy; Customer Focus; Workforce Focus (for HR Development & Management); Operations Focus (for Process Management); and Results (for Business Results).

Figure 2 Flynn and Saladin's (2001) Path Model Based on 1997 Baldrige Framework



A primary source of variation in model specification between studies has been the number of relationships hypothesized between the seven categorical CPE constructs. Number of proposed paths between constructs has varied from 10 (Ghosh et al., 2000) to 21 (Badri et al., 2006), with no two studies examining the same number of linkages. Once again, Figure 1 is a source of confusion. For example, Figure 1 joins the leadership, strategy-making, and customer focus constructs in a cluster with small two-headed arrows, and relates this cluster to other parts of the model with large two-headed arrows. Do the large arrows imply that the leadership, strategy, and customer constructs are each related to all of the other constructs, or are smaller numbers of more selective relationships involved? In some cases, researchers have liberally interpreted the number of linkages embedded in the CPE framework (e.g., Badri et al., 2006; He et al., 2011) while others have been more conservative (e.g., Ghosh et al., 2000; Lee et al., 2003).

Regardless of number of paths hypothesized, measurement models across all studies in Table 1 have generally exhibited acceptable overall levels of statistical fit and significance, thus prompting conclusions that empirical studies generally support the validity of the CPE framework (e.g., Bou-Llusar et al., 2009). However, investigations have generally been unable to confirm all hypothesized linkages, with non-significant paths often amounting to 20% or more of total paths proposed. Comparing results across studies reveals no common non-significant paths. Two studies that did find all paths significant (Badri et al., 2006; He et al., 2011) also happen to propose the most relationships between constructs. This is counterintuitive from a statistical standpoint, and suggests that issues associated with model development and testing in these studies may limit the generalizeability of the findings from these two studies.

MODEL AND HYPOTHESES

Rather than adding to the model proliferation of past studies, we reviewed the studies in Table 1 for a previously developed model of the CPE framework that could provide a useful basis for a follow-up empirical test, and selected Flynn & Saladin's (2001) model shown in Figure 2. We found this model attractive for a number of reasons. In addition to being firmly grounded in previous research, Flynn and Saladin's (2001) model is unique in that researchers were careful to conceptually link it to previous generations of CPE frameworks, thereby strengthening the logic of the researchers' causal model as an accurate representation of the nomological network intended by the CPE developers. Although three of the fifteen hypothesized paths between the seven CPE constructs were found to be non-significant, it is unclear as to whether this was due to possible model misspecification, the possibility of which the authors did not rule out in their discussion of results, or due to methodological limitations. Regardless, Flynn and Saladin's (2001) analysis of overall model fit and strength of measured relationships produced statistical results consistent with standards for well-performing measurement models. Flynn & Saladin's (2001) model is also attractive because it is a prominent one. It is well cited, and is used as a basis for discussion in several subsequent empirical studies listed in Table 1. As such, this model provides a solid foundation for a more decisive empirical test of the CPE framework's validity.

The Flynn and Saladin (2001) model (Figure 2) begins with leadership as the "driver." Leadership is proposed to directly influence downstream organizational process constructs related to information and analysis, strategy-making, and customer focus. Stated more formally using CPE category language:

Hypothesis 1a: Leadership has a direct, positive influence on Measurement, Information, and Analysis.
Hypothesis 1b: Leadership has a direct, positive influence on Strategic Planning.
Hypothesis 1c: Leadership has a direct, positive influence on Customer Focus.
Hypothesis 1d: Leadership has a direct, positive influence on Results.

The position of the information and analysis construct to the right of leadership in Flynn and Saladin's (2001) model reflects its broad influence on other organizational processes as depicted by the CPE developers in Figure 1. Information is proposed to directly influence thefour downstream organizational process constructs as well as the results construct:

Hypothesis 2a: Measurement, Information, and Analysis has a direct, positive influence on Strategic Planning.
Hypothesis 2b: Measurement, Information, and Analysis has a direct, positive influence on

Customer Focus.
Hypothesis 2c: Measurement, Information, and Analysis has a direct, positive influence on Workforce Focus.
Hypothesis 2d: Measurement, Information, and Analysis has a direct, positive influence on Operations Focus.
Hypothesis 2e: Measurement, Information, and Analysis has a direct, positive influence on Results.

The strategy making and customer focus constructs are positioned equidistantly and to the right of information and analysis. Strategy-making is proposed to influence customer focus as well as workforce focus. Customer focus is seen as influencing process management. Neither strategy-making nor customer focus processes are proposed to directly influence results:

Hypothesis 3a: Strategic Planning has a direct, positive influence on Customer Focus.
Hypothesis 3b: Strategic Planning has a direct, positive influence on Workforce Focus.
Hypothesis 4: Customer Focus has a direct, positive influence on Operations Focus.

The final two organizational processes related to workforce focus and process management are positioned equidistantly and to the right of strategy making and customer focus. These two processes also directly precede the results construct in the nomological network. Workforce focus is proposed to influence both process management and downstream results. Process management, being the operational instrument of all other organizational process activities, is proposed to influence only downstream results:

Hypothesis 5a: Workforce Focus has a direct, positive influence on Operations Focus.
Hypothesis 5b: Workforce Focus has a direct, positive influence on Results.
Hypothesis 6: Operations Focus has a direct, positive influence on Results.

Since the framework's establishment, a central proposition of the CPE is that it is applicable to organizations regardless of sector (Bell & Keys, 1998; Douglas & Fredendall, 2004). As suggested by Table 1 above, empirically testing the robustness of this proposition has been problematic due to the inherent difficulty of obtaining cross industry samples that adequately represent, for example, the manufacturing, service, healthcare, education, and non-profit sectors for which the CPE have been categorically developed. The literature provides conflicting perspectives regarding the applicability of Baldrige framework across industries. Studies of total quality management frameworks have sometimes suggested universal orientation across product-based and service-based sectors (e.g., Douglas & Fredendall, 2004; Prajogo, 2005). On the other hand, many studies highlight the differences between quality management across industries, particularly between manufacturing and service organizations (e.g., Huq & Stolen, 1998; Ronnback & Witell, 2008; Woon, 2000). To investigate this issue in the context of this study, we formulate the following hypothesis:

Hypothesis 7: The same Baldrige path model is applicable across the various categorical sectors of the CPE.

DATA AND METHOD

Data

The Baldrige Award review process consists of three stages. The first stage is the Independent Review, during which all applicants gets itemized numerical scores from a panel of examiners working independently based on the written application. Selected applicants will then advance to the second stage, the Consensus Review. At this stage the examiner team will discuss variation in the individual scoring process and reach consensus on the itemized scores. Lastly, the examiner teams visit highest-scoring applicants to interview personnel and verify information contained in written applications. During the site visit review only site visit reports are written but the applicants are not scored.

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. 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).

The Baldrige evaluation process at the national level 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. The Baldrige standard has been adopted in the U.S as well as around the world. The former director of the Baldrige program has suggested that the greatest success of the award is in the area of assessment (Bell and Keys 1998). Evans and Mai (2014) reported that using confirmatory factor analysis on the NIST applicant scoring dataset, there is strong evidence of reliability, convergent and discriminant validity in the Baldrige framework, which is consistent with other research that used survey or state-level data. Compared with state-level data and self-reported surveys, the use of objective and reliable applicant data allows us to make more generalizable conclusions.

There are several hurdles that discourage us from analyzing the released itemized scoring data set directly. First of all, the Baldrige framework has gone through several revisions since 1988, most notably in 1992 and 1997. In these revisions, new items have been added in the framework and the conceptual relationships between the core categories have been reorganized. Flynn and Saladin (2001) demonstrate that the newer framework offers a better fit for empirical survey data and therefore concluded that the revisions made to the Baldrige framework are beneficial. Second, trends can be observed in the applicant scoring data for several industry sectors. For instance, manufacturing applicants' scores in many categories show a declining trend during the years in the data set. Hence, analyzing the relationship between the categories without considering the time trends can produce spurious results.

In light of these challenges, we used the Independent Review scores from the applicants after 1999 in our study. The main reason for only including applicants in post-1999 period is that no significant change has been made to the framework. In addition, the number of Baldrige scoring items were abridged after 1999, each corresponding to a clear and well-defined construct under a category. Each item's raw score was converted to a percentage score, which was done by dividing it by the maximum possible score for that item. Only a small number of aggregations was needed for consistency. For instance, in Baldrige Criteria after 2003, item 6.2 and 6.3 were combined to a single item 6.2 to measure the relationship of work processes to work systems. We aggregated items 6.2 and 6.3 from 1999-2002 in a similar way. Lastly, we adjusted each item score by subtracting the mean item score of that year. Adjusting the yearly mean for each item can de-trend the time series and make the spurious regression less likely. We were left with 500 observations across five industry sectors. Table 2 shows the industry sector distribution of our sample.

Table 2: Industry Distribution of Sample

Industry Sector	***N***	**Percent (%)**
Manufacturing	78	15.6%
Service	58	11.6%
Small Business	103	20.6%
Education	99	19.8%
Healthcare	152	30.4%
Nonprofit	10	2.0%

Method

As discussed above, we adopt a structural model similar to Flynn and Saladin (2001) for our analysis (Figure 2). We used partial least squares (PLS) structural equation modeling to analyze the data. Traditionally, hypotheses regarding the Baldrige framework have been tested using covariance-based SEM (CBSEM). Our rationale for using PLS is as follows. First, although the overall sample size (N = 500) is rather large, a causal model built upon the Baldrige framework combined with all the measurement items can have many parameters. When we make comparison across industry sectors the statistical power of CBSEM can be affected, since the sample size within each group is less than the rule of thumb of 5 observations per parameter estimated (Shah & Goldstein, 2006). PLS is more flexible in this regard. It only requires a sample size of 10 times the most complex relationship in the research model (Peng & Lai, 2012). Second, many of theoretical models on quality management involve relative large number of paths between categories. PLS uses iterative method to estimates the factor loadings and structural paths in models and therefore is more appropriate to assess the predictive power in complex models (Peng & Lai, 2012). Third, many of the items scores are non-normal and PLS provides more robust estimates in the presence of non-normality (Cassel et al., 1999). We use *plspm* library in R 2.13.1 to estimate our research model.

MODEL EVALUATION AND RESULTS

Measurement Model

The measurement quality of our PLS structural equation model can be validated through internal reliability, convergent, and discriminant validities. Table 3 shows that all measurement item loadings are greater than 0.70, indicating convergent validity at indicator level. The convergent validity is also satisfactory at construct level since the average variance extracted (AVE) values are greater than 0.50 (Peng & Lai, 2012). The composite reliabilities are greater than 0.70, exceeding the recommended values (Hair et al., 2011). Evidence of discriminant validity is obtained by examining the diagonal elements (the square root of AVE) of the correlation matrix shown in Table 4. The square root of the average variance explained by each latent variable is greater than the correlations between that latent variable and other latent variables in the model, demonstrating discriminant validity (Chin, 1998). In addition, we also confirmed that an item's loading with its associated latent construct is higher than its loadings with all the remaining constructs (i.e., the cross loadings) (Hair et al., 2011). Collectively, the evidence suggests that the Baldrige constructs have adequate measurement properties.

Table 3 Measurement Properties of Baldrige Categories

Construct	**Item**	**Item loading**	**Composite reliability**	**AVE**
Leadership	Leadership approach and deployment (1.1)	0.948	0.941	0.888
	Societal Responsibility (1.2)	0.936		
Strategic Planning	Strategy development (2.1)	0.963	0.920	0.926
	Strategy deployment (2.2)	0.961		
Measurement, Information, and Analysis	Measurement and analysis (4.1)	0.942	0.936	0.880
	Information Management (4.2)	0.934		
Customer Focus	Customer and market knowledge (3.1)	0.951	0.947	0.898
	Customer relationship building and satisfaction determination (3.2)	0.944		
Workforce Focus	Work Systems (5.1)	0.894	0.872	0.797
	Education, Training, and Development (5.2)	0.899		
	Employee well-being and satisfaction (5.3)	0.885		
Operations Focus	Value creation processes (6.1)	0.943	0.881	0.894
	Support processes (incl. business and supplier) (6.2)	0.947		
Results	Customer and product results (7.1, 7.2)	0.909	0.932	0.831
	Process/operational results (7.5)	0.929		
	Financial and market results (7.3)	0.896		
	Human resource results (7.4)	0.912		

Table 4 Correlations of Latent Constructs

	L	SP	MIA	CF	WF	OF	R
L	**0.942**						
SP	0.826	**0.962**					
MIA	0.847	0.831	**0.938**				
CF	0.830	0.804	0.790	**0.948**			
WF	0.812	0.764	0.761	0.777	**0.893**		
OF	0.801	0.820	0.797	0.775	0.736	**0.946**	
R	0.792	0.772	0.790	0.762	0.746	0.762	**0.912**

Note: Values on diagonal are the square root of construct's average variance extracted (AVE)

Structural Model

We first used the entire data set to fit the model using PLS analysis. The overall quality of the research model can be assessed using Goodness of Fit (GoF) following Tenenhaus et al., (2005). The Flynn and Saladin (2001) model result in a Gof of 0.782, which is well above the suggested cut-off value of 0.36. The Stone-Geisser's Q^2 (Geisser, 1975; Stone, 1974) was used to assess the predictive relevance for endogenous variables using a blindfold procedure. With Q^2 greater than 0 (ranging from 0.42 to 0.64) for all the six endogenous variables, we can conclude that the model has considerable predictive power.

Table 5 reports the path coefficients of the CPE model estimated for the entire data set. The explanatory power of the structural model can be evaluated by looking at the R^2 value in the dependent constructs. The R^2 values for the CPE constructs are reported in Table 6. With fairly high values for all endogenous variables, especially with the R^2 being 0.699 for the Results category, we can conclude that the model has substantial explanatory power for quality management actions and outcomes. The 95% confidence intervals are estimated using 500 bootstrapped samples. We confirmed the consistency of this result by running the structural model's bootstrap procedure with 200 and 1000 times of resampling.

Table 5 Structural Estimates of PLS analysis

Path	**PLS result**		
	Coefficient	**Std. Error**	**95% CI**
Leadership -> Measurement, Information, and Analysis	0.825	0.0131	(0.800, 0.849)
Leadership -> Strategic Planning	0.506	0.0447	(0.412, 0.589)
Leadership -> Customer Focus	0.438	0.0537	(0.332, 0.541)
Leadership -> Results	0.294	0.0518	(0.192, 0.389)
Measurement, Information, and Analysis -> Strategic Planning	0.414	0.0426	(0.334, 0.497)
Measurement, Information, and Analysis -> Customer Focus	0.304	0.0433	(0.222, 0.389)
Measurement, Information, and Analysis -> Workforce Focus	0.424	0.0478	(0.324, 0.516)
Measurement, Information, and Analysis -> Operations Focus	0.492	0.0395	(0.412, 0.565)
Measurement, Information, and Analysis -> Results	0.204	0.0509	(0.103, 0.297)
Strategic Planning -> Customer Focus	0.167	0.0539	(0.067, 0.268)
Strategic Planning -> Workforce Focus	0.409	0.0484	(0.319, 0.511)
Customer Focus -> Operations Focus	0.252	0.0445	(0.165, 0.339)
Workforce Focus -> Operations Focus	0.164	0.0385	(0.088, 0.240)
Workforce Focus -> Results	0.191	0.0421	(0.109, 0.273)
Operations Focus -> Results	0.218	0.0505	(0.118, 0.314)

Table 6 R^2 for Endogenous Variables in PLS Model

Endogenous variables	**R^2**	**95% CI for R^2**
Strategic Planning	0.681	(0.640, 0.721)
Measurement, Information, and Analysis	0.772	(0.734, 0.804)
Customer Focus	0.739	(0.695, 0.776)
Workforce Focus	0.635	(0.588, 0.684)
Operations Focus	0.720	(0.673, 0.758)
Results	0.699	(0.658, 0.740)

We tested H1a - H1d by examining the structural model's coefficients for Leadership to other constructs. As seen in Table 5, the coefficients vary from 0.294 0.825, all of which are significant at $\alpha = 0.05$ level. The results support the hypothesis H1a - H1d that senior leadership should have direct, positive impact on Measurement, Information and Analysis, Strategic Planning, Customer Focus, and Results. The magnitude of the direct linkage coefficient from Leadership to Results, however, is not strong (0.294). This suggests that results of the performance excellence system hinge on strong commitments from other parts of the organization besides leadership.

To test H2a - H2e, we first note that in the Flynn and Saladin (2001) model, Measurement, Information, and Analysis serves as a critical role in the system. It is considered the most important endogenous variable in the causal model and is the linkage between management effort and other components in the system. The PLS analysis confirms this role, as its path coefficients to Strategic Planning, Customer Focus, Workforce Focus, Operations Focus and Results are all positive and statistically significant. In addition, we fit a PLS model without the Measurement, Information and Analysis construct included. The reduced model has smaller goodness of fit measure and other endogenous variables have Stone-Geisser's Q^2 measures. This indicates that an organization's performance in Measurement, Information and Analysis is important in predicting the overall quality management outcome. Overall the evidence supports H2a - H2e.

To test H3a, H3b, and H4, we examined the structural model's path coefficients from Strategic Planning to Customer Focus and Workforce Focus, as well the path coefficient from Customer Focus to Operations Focus. Once again, the path coefficients are all positive and statistically significant. The magnitude of the coefficient between Strategic Planning and Customer Focus, estimated at 0.167, is relatively low, suggesting a reduced role of strategic planning on customer and market processes compared to other categorical constructs such as those related to leadership and information and analysis. Overall, however, the evidence supports H3a, H3b, and H4.

Table 7 Structural Estimates by Industry Sector

	Manufacturing		**Service**		**Small Business**		**Education**		**Healthcare**	
	Coefficient	Std. Error	Coefficient	Std. Error	Coefficient	Std. Error	Coefficient	Std. Error	Coefficient	Std. Error
L -> MIA	**0.77***	0.05	**0.89***	0.03	**0.85***	0.02	**0.80***	0.03	**0.80***	0.02
L -> SP	**0.63***	0.08	**0.57***	0.12	**0.38***	0.11	**0.43***	0.07	**0.59***	0.08
L -> CF	**0.46***	0.14	**0.41***	0.19	**0.36***	0.11	**0.51***	0.10	**0.42***	0.09
L -> R	0.03	0.12	0.27	0.20	**0.47***	0.10	0.17	0.13	**0.32***	0.11
MIA -> SP	**0.27***	0.09	**0.36***	0.11	**0.53***	0.10	**0.51***	0.07	**0.30***	0.08
MIA -> CF	**0.28***	0.09	0.00	0.13	**0.50***	0.11	**0.30***	0.09	**0.24***	0.09
MIA -> WF	**0.33***	0.14	0.27	0.17	**0.51***	0.10	**0.28***	0.12	**0.43***	0.08
MIA -> OF	**0.44***	0.11	**0.51***	0.10	**0.50***	0.09	**0.50***	0.10	**0.38***	0.09
MIA -> R	0.20	0.13	0.02	0.13	**0.33***	0.11	0.18	0.11	**0.21***	0.09
SP -> CF	0.14	0.13	**0.47***	0.14	0.05	0.14	0.13	0.11	**0.24***	0.09
SP -> WF	**0.37***	0.12	**0.57***	0.16	**0.38***	0.10	**0.57***	0.11	**0.38***	0.09
CF -> OF	**0.38***	0.11	**0.32***	0.13	**0.26***	0.09	0.18	0.11	**0.19***	0.09
WF -> OF	0.08	0.09	0.09	0.12	0.15	0.09	**0.25***	0.09	**0.27***	0.07
WF -> R	**0.36***	0.09	**0.27***	0.14	-0.04	0.08	**0.17***	0.09	**0.24***	0.08
OF -> R	**0.34***	0.12	**0.35***	0.14	**0.18***	0.09	**0.37***	0.12	0.09	0.10

*****: significant at the $p < 0.05$ level

To test H5a, H5b, and H6, the path coefficient from Workforce Focus to Operations Focus was examined, as well as the coefficients from Workforce Focus and Operations Focus to Results. All path coefficients are positive and significant at the $\alpha = 0.05$ level. However, ranging from 0.164 to 0.218, the magnitudes of these coefficients are among the weakest in the model. These results are consistent with Flynn and Saladin's (2001) analysis that found generally weakening relationships in the model's downstream paths involving Workforce Focus, Process Management, and Results. Evidence presented here supports H5a, H5b, and H6 nonetheless.

Lastly, we conducted multiple-group analyses to examine whether the relationships between the categories are moderated by industries (H7). We omitted the nonprofit sector due to its small sample size in the data set. We conducted the group comparison at both the measurement level and structural level. We first confirmed that the measurement model is indeed adequate and similar for all groups. Then we focused on the structural level differences. That is, we wanted to test whether the magnitude of the structural coefficients (i.e., the path coefficients) are the same for all industry sectors. We used the Chin/Keil approach (Keil et al., 2000) to conduct the analysis. More specifically, for each for the industry sectors the path coefficients were estimated separately, and the standard errors and non-parametric confidence intervals of the coefficients were computed from bootstrapping observations in each group. Table 7 displays the group-specific path coefficient estimates. H7 is clearly rejected, since for all but the first three paths only a subset of industry sectors have significant coefficients.

DISCUSSION

The primary objective of this investigation was to provide a more decisive test of the CPE framework by addressing several shortcomings of previous studies. Use of Flynn and Saladin's (2001) representation of the third generation CPE framework offered a prominent, well-cited model firmly anchored in theory that provided a basis for comparative evaluation. Use of Baldrige applicant scoring data offered both a large, cross sectional sample pool and a primary data source of expert observers with extensive skill in assessing organizational activities in light of the CPE framework to overcome methodological limitations of past studies.

Findings from our analysis stemming from this approach found strong support for the third generation CPE framework as modeled by Flynn and Saladin (2001). When the entire 500 sample cross industry data set was used to estimate the structural model pictured in Figure 2, all hypothesized relationships between CPE constructs were found significant. Similar to Flynn and Saladin's (2001) analysis, the strength of those relationships as reflected by the structural model's path coefficients was found to decline among the "downstream" constructs of Workforce Focus, Process Focus, and Business Results. When discussing their results, Flynn and Saladin (2001) suggested that this might reflect failure to account for some important relationships not included in the third generation CPE framework. However, several of the relationships found to be non-significant in Flynn and Saladin's (2001) model were found significant here, suggesting that methodological differences such as sample size and composition—issues that Flynn and Saladin (2001) discussed as potential limitations of their study—may help explain findings in this context. Providing a more favorable backdrop for evaluating the third generation CPE framework's general validity, as well as associated evidence that supports the framework in its entirety, constitute the primary contributions of this study.

Although analysis based on the entire cross industry sample afforded strong support for the CPE framework in its entirety, that support did not transfer to analysis of the framework on a sector-by-sector basis. As Table 7 indicates, path coefficients between CPE constructs were found non-significant in all sector models, ranging from 1 in 15 non-significant for the health care sector structural model to 5 in 15 non-significant for the service sector structural model. Differences in sector sample size might contribute to these results. Matching the sector sample sizes appearing in Table 2 to the number of non-significant paths in Table 7 does suggest some degree of relationship. Because overcoming small, homogenous sector samples of previous studies was a primary methodological limitation that the present investigation sought to overcome when evaluating the CPE framework's general validity, then perhaps the findings in Table 7 should not be surprising. However, explanations in terms of sample issues alone are not completely satisfactory. As shown in Table 7, only 4 of the 15 path coefficients were significant across all sectors, and no discernable pattern of non-significance is evident. Moreover, low values associated with many of the non-significant path coefficients suggest lack of relationship between particular constructs regardless of sample size. Coupling historical arguments for sector-dependent quality system designs (e.g., Huq and Stolen 1998; Ronnback and Witell 2008; Woon 2000) with results reported here raises questions as to whether the performance excellence frameworks are structured differently between sectors. Investigations that assume a contingency-based stance to explore such inter-sector differences seems an interesting future research opportunity.

REFERENCES

Badri, M.A., Selim, H., Alshare, K., Grandon, E.E., Younis, H., & Abdulla, M. (2006). The Baldrige Education Criteria for performance excellence framework: Empirical test and validation. *International Journal of Quality & Reliability Management, 23,* 1118-1157.

Bell, R. & Keys, B. (1998). A conversation with Curt W. Reimann on the background and future of the Baldrige award. *Organizational Dynamics, 26*(4), 51-61.

Bemowski, K. & Stratton, B. (1995). How do people use the Baldrige award criteria? *Quality Progress, 28*(5), 43-48.

Bou-Llusar, J.C., Escrig-Tena, A.B., Roca-Puig, V., & Beltrán-Martín, I. (2009). An empirical assessment of the EFQM Excellence Model: Evaluation as a TQM framework relative to the MBNQA Model. *Journal of Operations Management, 27*, 1-22.

Cassel, C., Hackl, P., & Westlund, A.H. (1999). Robustness of partial least-squares method for estimating latent variable quality structures. *Journal of Applied Statistics,26*, 435-446.

Chin, W.W. (1998). The partial least squares approach to structural equation modeling. In G.A. Marcoulides (ed.), *Modern methods for business research*, Hillsdale, NJ: Lawrence Earlbaum Associates, 294-336.

Cronbach, L.J., & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281-302.

Curkovic, S., Melnyk, S., Calantone, R., & Handfield, R. (2000). Validating the Malcolm Baldrige National Quality Award framework through structural equation modelling. *International Journal of Production Research, 38*, 765-791.

Dearborn, D.C., & Simon, H.A. (1958). Selective perception: A note on the departmental identifications of executives. *Sociometry, 21*, 140-144.

Douglas, T.J., & Fredendall, L.D. (2004). Evaluating the Deming management model of total quality in services. *Decision Sciences, 35*, 393-422.

Ettorre, B. (1996). Is the Baldrige still meaningful? *Management Review, 85*, 28-43.

Evans, J.R, Ford, M.W., Masterson, S.S., & Hertz H.S. (2012). Beyond performance excellence: research insights from Baldrige recipient feedback. *Total Quality Management & Business Excellence, 23*, 489-506.

Evans, J.R., & Mai, F. (2014). Insights from the Baldrige Award item-level blinded applicant scoring data. *Quality Management Journal, 21*(2), 45-64.

Flynn, B.B., & Saladin, B. (2001). Further evidence on the validity of the theoretical models underlying the Baldrige criteria. *Journal of Operations Management, 19*, 617-652.

Flynn, B.B., & Saladin, B. (2006). Relevance of Baldrige constructs in an international context: a study of national culture. *Journal of Operations Management, 24*, 583-603.

Ford, M.W., & Evans, J.R. (2001). Baldrige assessment and organizational learning: The need for change management. *Quality Management Journal, 8*(3), 6-22.

Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association, 70*, 320-328.

Ghosh, S., Handfield, R.B., Kannan, V.R., & Tan, K.C. (2003). A structural model analysis of the Malcolm Baldrige National Quality Award framework. *International Journal of Management and Decision Making, 4*(4), 289-311.

Goldstein, S.M., & Schweikhart, S.B. (2002). Empirical support for the Baldrige Award framework in US hospitals. *Health Care Management Review, 27*(1), 62-75.

Hair, J.F., Ringle, C.M., & Sarstedt, M. (2011). PLS-SEM: Indeed a silver bullet. *Journal of Marketing Theory and Practice, 19*(2), 139-52.

Hambrick, D.C. (1981). Strategic awareness within top management teams. *Strategic Management Journal, 13*, 145-158.

He, Z., Hill, J., Wang, P., & Yue, G. (2011). Validation of the theoretical model underlying the Baldrige criteria: Evidence from China. *Total Quality Management, 22*(2), 243-263.

Huq, Z., & Stolen, J.D. (1998). Total quality management contrasts in manufacturing and service industries. *International Journal of Quality & Reliability Management, 15*, 138-161.

Jayamaha, N.P, Grigg, N.P., & Mann, R.S. (2008). Empirical validity of Baldrige criteria: New Zealand evidence. *International Journal of Quality & Reliability Management, 25*, 477-93.

Jayamaha, N.P., Palitha, N., Grigg, N.P., & Mann, R.S. (2011). Empirical analysis of the Baldrige Criteria as both an organisational performance measure and a theoretical model. *Measuring Business Excellence, 15*(1), 20-33.

Karimi, A., Safari, H., Hashemi, S.H., and Kalantar, P. (2014). A study of the Baldrige Award framework using the applicant scoring data. *Total Quality Management, 25*, 461-477.

Keil, M., Tan, B.C.Y., Wei, K.K., Saarinen, T., Tuunainen, V. & Wassenaar, A. (2000). A cross-cultural study on escalation of commitment behavior in software projects. *MIS Quarterly, 24*, 299-325.

Lee, S.M., Rho, B.H., & Lee, S.G. (2003). Impact of Malcolm Baldrige National Quality Award Criteria on organizational quality performance. *International Journal of Production Research, 41*, 2003-2020.

Meyer, S.M., & Collier, D.A. (2001). An empirical test of the causal relationships in the Baldrige Health Care Pilot Criteria. *Journal of Operations Management, 19*, 403-425.

Nisbett, R.E., & Wilson, T.D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*(3), 231-259.

Pannirselvam, G.P. & Ferguson, L.A. (2001). A study of the relationships between the Baldrige categories. *International Journal of Quality & Reliability Management, 18*, 14-37.

Peng, D.X., & Lai, F. (2012). Using partial least squares in operations management research: A practical guideline and summary of past research. *Journal of Operations Management, 30*, 467-480.

Prajogo, D.I. (2005). The comparative analysis of TQM practices and quality performance between manufacturing and service firms. *International Journal of Service Industry Management, 16*(3), 217-228.

Ronnback, A., & Witell, L. (2008). A review of empirical investigations comparing quality initiatives in manufacturing and service organizations. *Managing Service Quality, 18*, 577-93.

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

Steeples, M.M. (1992). *The corporate guide to the Malcolm Baldrige national quality award: Proven strategies for building quality into your organization*. Milwaukee: ASQ Quality Press.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, 36*(2), 111-147.

Tenenhaus, M., Vinzi, V.E., Chatelin, Y.M., & Lauro, C. (2005). PLS path modeling. *Computational Statistics & Data Analysis, 48*(1), 159-205.

Wilson, D.D., & Collier, D.A. (2000). An empirical investigation of the Malcolm Baldrige National Quality Award causal model. *Decision Sciences, 31*, 361-383.

Woon, K.C. (2000). TQM implementation: Comparing Singapore's service and manufacturing leaders. *Managing Service Quality, 10*(5), 318-331.