Empirical analysis of the Baldrige Criteria as both an organisational performance measure and a theoretical model

Nihal Palitha Jayamaha, Nigel Peter Grigg and Robin Stephen Mann

Nihal Palitha Jayamaha is a Lecturer and Nigel Peter Grigg is an Associate Professor, both at the Centre for Industrial Management and Innovation, School of Engineering and Advanced Technology, Massey University, New Zealand. Robin Stephen Mann is Director of the Centre for Organisational Excellence Research, School of Engineering and Advanced Technology, Massey University, New Zealand.

Summary

Purpose – *The purpose of this paper is to test empirically two key measurement perspectives – measurements in the context of a theoretical model that predicts/explains results, and measurements in the context of generating an overall score on performance excellence – of the Baldrige Criteria for Performance Excellence (BCPE) using data from Australasian Business Excellence Award applicants.*

Design/methodology/approach – *Two theoretical models corresponding to each measurement perspective were tested using data (n = 118) from the applicants for the New Zealand Businesses Excellence Award (based on the BCPE). The partial least squares method was used to test the validity of the measurement items of the BCPE. Qualitative data were also collected from applicants for the Australian Business Excellence Award.*

Findings – *Most of the measurement items showed low levels of measurement validity under both measurement perspectives; the main reason for this was considered to be due to the design of the BCPE as it is a high level of integration and alignment between the various items and categories.*

Research limitations/implications – *The study is based on a non-probability sample, although this was unavoidable because the desire was to use data on national business excellence award applicants; such data are rarely available for research purposes.*

Practical implications – *The paper shows that it is important for organisations to understand the relationships between the various categories/items. Also, the paper indicates that more research should be undertaken in this area to assist organisations to understand the relationships.*

Originality/value – *The dual measurement perspectives of a BE model (using the same data) has not been tested before, and it is hoped that this study will help academia and the practitioner community to develop more refined performance excellence measures.*

Keywords *Baldrige Award, Non-profit organizations, Least square approximation, Australasia*

Paper type *Research paper*

1. Introduction

Business excellence (BE) models, particularly the Baldrige Criteria for Performance Excellence (BCPE) and the European Foundation of Quality Management (EFQM) Excellence Model, are being used by organisations all over the world for self-assessment, benchmarking, sharing of best practices and assessing organisations for national quality/BE awards (Dahlgaard *et al.*, 1998; Porter and Tanner, 2004). As an assessment tool, a BE model assesses the achievements of an organisation in specific areas – known as categories, which are in turn divided into various measurement items – against stipulated assessment criteria. The most rigorous forms of assessments take place when organisations are assessed by trained examiners under national quality/BE award schemes (Garvin, 1991; Jayamaha *et al.*, 2009; Pannirselvam *et al.*, 1998).

The measurement items of a BE model should fulfil two theoretical requirements. First, they should adequately represent their respective constructs (categories), which in turn should represent a theoretical model that predicts and explains results (Figure 1). At the same time

Figure 1 BCPE framework

Source: NIST (2009)

items should (via their respective categories) collectively constitute an overall performance excellence index or score that can be reliably used for measuring overall performance achievement of an organisation, for the purpose of self-assessment, benchmarking and selecting winners at national quality awards (Dean and Tomovic, 2004; Williams *et al.*, 2006). Thus, BE models should be both theoretically valid and practically useful to organisations. In this study, using the BCPE as a frame of reference and based upon empirical data resulting from national quality award applications in New Zealand and Australia, we present an analysis of the validity of a BE model from each of these theoretical perspectives.

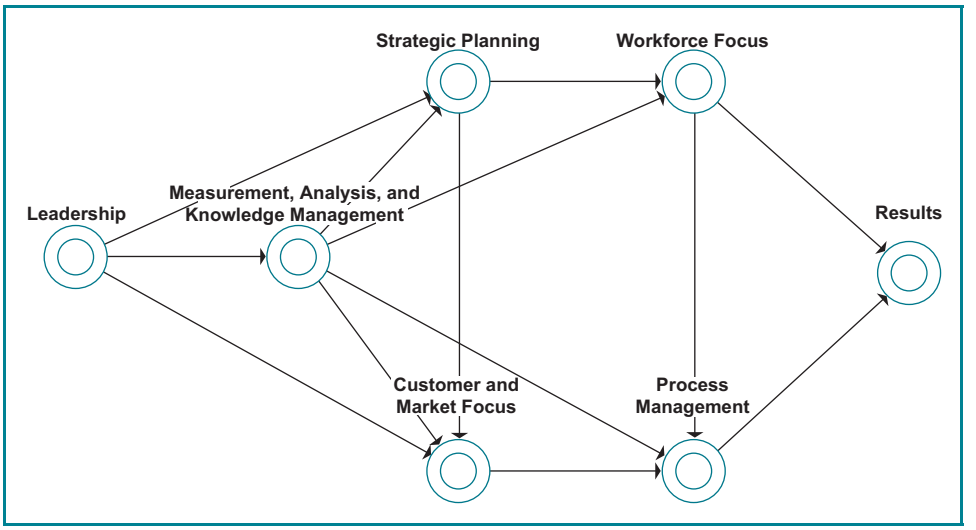
2. Literature review: the empirical validation of business excellence models

The BCPE, as a seven-construct framework (Figure 1), explains how results are achieved given the environment in which an organisation operates. As evidenced from Figure 1, the BCPE framework involves many bi-directional paths between constructs, which makes structural equation modeling extremely complex. Hence researchers have used parsimonious models to test the validity of the BCPE (Collier *et al.*, 2002). The parsimonious theoretical model of the BCPE (Figure 2) implies that results are achieved by putting in place high performance work systems and other human motivational practices coupled with efficient and effective processes; these in turn being influenced by the senior leadership, through creating a long-term focus on customers and markets though data and knowledge-based decision making (Flynn and Saladin, 2001: Pannirselvam and Ferguson, 2001).

2.1 Empirical validations involving the BCPE Model

Validity of the BCPE as a theoretical model has been examined by many past researchers, typically using proxy instruments (e.g. questionnaires) to obtain measurements that capture the essence of the seven constructs (data based on the actual test instrument or test conditions are not available due to strict confidentiality requirements imposed by the bodies responsible for administering these models – in the USA the National Institute of Standards and Technology (NIST) administers the Malcolm Baldrige National Quality Award). While the measurement items of the seven constructs of the BCPE are revised annually (often resulting in minor changes) by NIST, as a theoretical model, the BCPE has remained fairly stable since

Figure 2 The parsimonious BCPE theoretical model (Model I)



1997. However, the BCPE continues to evolve from a model of manufacturing quality (pre-1997 versions) towards becoming a more universal model of overall organisational performance excellence (National Institute of Standards and Technology, 2009). As a major departure from the current thinking on BE, we note that in pre-1997 BCPE models, customer satisfaction was posited to be the final outcome variable (as opposed to key stakeholder outcomes in post-1997 BCPE models), which was posited to be caused by achievement of quality assurance (QA) goals; this conceptualisation is consistent with manufacturing quality thinking (Anderson *et al.*, 1994; Deming, 1986). We therefore focus our attention mostly on studies on post-1997 versions of the BCPE and equivalent models.

Based on data ($n = 164$) collected from manufacturing plants of six industrialised countries, Flynn and Saladin (2001) compared the goodness-of-fit measures of the 1997 BCPE theoretical model against two prior versions, using the path modelling technique, which is a basic form of covariance based structural equation modelling (CBSEM). They found that the goodness-of-fit of the 1997 model was superior to those corresponding to the 1988 and 1992 versions. Hence they argued that conceptualisation of the theoretical constructs of the BCPE is progressing in the right direction. Based on data collected from USA hospitals ($n = 220$), Meyer and Collier (2001) showed that the BCPE theoretical model fits data reasonably satisfactorily with a root mean square error of approximation (RMSEA) of 0.086, although both the ''Process Management'' and ''Workforce Focus'' constructs were found to have no direct relationship with results. This finding to a certain extent invalidates the underlying theory represented by post-1997 models, in which Process Management and Workforce Focus are deemed to be the causal antecedents of organisational outcomes (results). Lee *et al.* (2003) used CBSEM to study the validity of the BCPE theoretical model using data ($n = 109$) collected from Korean manufacturing plants. However, they used only measures on quality performance to capture the conceptual domain of the ''results'' construct – whereas quality performance is only one aspect of organisational outcomes. Lee *et al.* (2003) found that their model was marginally acceptable in terms of goodness-of-fit to data.

Using the Arizona Governor's Quality award as a proxy for BCPE, Pannirselvam and Ferguson (2001) studied the validity of the BCPE theoretical model. They used individual examiner scores ($n = 272$) of the measurement items of the 67 organisations that applied for the aforesaid state quality award in 1993. Using CBSEM, they found that their model was a reasonable fit to data with many of the structural paths of their theoretical model remaining statistically significant. Pannirselvam *et al.* (1998) also used the same data set to investigate the validity of the measurement items of the BCPE more closely. They found that the

measurement items of the BCPE possess construct validity based on their confirmatory factor analysis (CFA) model (CFA is a form of CBSEM).

Jayamaha *et al.* (2008) used partial least squares based structural equation modelling (PLSBSEM) to study the validity of the BCPE theoretical model based on data ($n = 91$) collected from New Zealand organisations using a self-assessment instrument. PLSBSEM is a piece-wise (partially) optimised multiple regression approach that is often used as an alternative to the more widely used CBSEM approach, which involves a global optimisation parameter (Chin, 1998). Jayamaha *et al.* (2009) also conducted a similar study based on data (item scores) collected from the applicants for national BE awards in three countries:

1. Australia;
2. New Zealand; and
3. Singapore.

In both studies the validity of the measurement items was found to be low, coming close to failing even the most basic requirements for measurement validity. This may lead one to speculate whether the measurement items of the BCPE in fact measure the underlying constructs satisfactorily. However, to our knowledge, no one has investigated whether or not the measurement items of the BCPE reflect their underlying constructs which in turn form an overall performance excellence index or score.

Considering the fact that the overall BE score given to an organisation is treated as the sum of the seven category scores (nine category scores in the case of EFQM Excellence Model) and that category scores themselves are sums of their corresponding item scores, it make sense to conceptualise the categories (constructs) as if they belong to a higher order construct (or concept), which may be called ''Business Excellence''.

Curkovic *et al.* (2000) conceptualised BE as a second-order construct formed by the seven categories of the BCPE (1997 version). However, their second-order construct had only four first-order constructs:

1. Total Quality Management (TQM) *Strategic Systems* (consisting of three BCPE categories, i.e. Leadership, Strategic Planning, and Customer and Market Focus);
2. TQM *Operational Systems* (consisting of two BCPE categories, i.e. Human Resource Focus, which is now known as Workforce Focus, and Process Management);
3. TQM *Information Systems* (consisting of the BCPE category Information and Analysis, which is now known as Measurement Analysis, and Knowledge Management); and
4. TQM *Results* (consisting of the BCPE category Business Results, which is now known as Results).

They used the CFM approach to test their second-order factor model using data from a sample of plant managers in the US automotive industry ($n = 526$). Curkovic *et al.* (2000) found that their factor model was a good fit to data, thus confirming that the higher order concept BE does exist.

2.2 Empirical validations involving the EFQM Excellence Model

Eskildsen *et al.* (2001) used a CFA model conceptualising BE as a construct that reflects the nine categories of the EFQM excellence model. They used data from a sample of Danish companies ($n = 756$) using a questionnaire designed to measure the nine categories of the EFQM Excellence Model. Having observed that their factor model was a moderately acceptable fit to data (RMSEA slightly higher than 0.1, which is the upper bound cut-off value for an acceptable fit), they went on to estimate the category weights using the factor scores generated by a component-based approximation. Based on the empirical category weights they determined, they concluded that the five enabler categories – that is the categories that represent what managers do to achieve results – are considerably more important (700 points) than the four results categories (300 points) for the Danish companies, thus asserting that the stipulated category weights of the EFQM Excellence

Model (500 points for enabler categories and 500 points for the results categories) are not universally valid. Using longitudinal data (1998-2001) Eskildsen *et al.* (2002) also showed that the category weights do not remain stable over time.

Bou-Llusar *et al.* (2005) conducted canonical correlation analysis (CCA) incorporated in CBSEM to test the relationship between the five enabler categories (conceptualised to form a single construct) and the four results categories (again, conceptualised to form a single construct) of the EFQM Excellence Model; they then reduced the two constructs, which form the ''canonical correlation pair'', to a single latent construct using the CCA procedures prescribed in the literature. Their model was tested with data from a sample of Spanish companies ($n = 446$). They found a strong linkage between the enabler categories and results categories. However they were not able to find a statistically significant difference between their model and the null model (which assumed equal weights for enabler and results categories, as stipulated in the EFQM Excellence Model), which implies that the stipulated weights of the EFQM Model are appropriate for the Spanish industry. There are also other researchers who have conceptualised the enabler categories as a single construct to study the linkage between enablers and results. For example, Prajogo and Brown (2004) conceptualised the enabler categories of the BCPE (i.e. all but the results category) as reflective of a single construct; they labelled this construct as TQM, which was hypothesised to be causally related to the results category. Likewise, Tamimi (1998) conceptualised the underlying constructs of the Deming Management Method (DMM) to be reflective of a second-order construct labelled TQM.

2.3 Gaps in existing research and the need for the present study

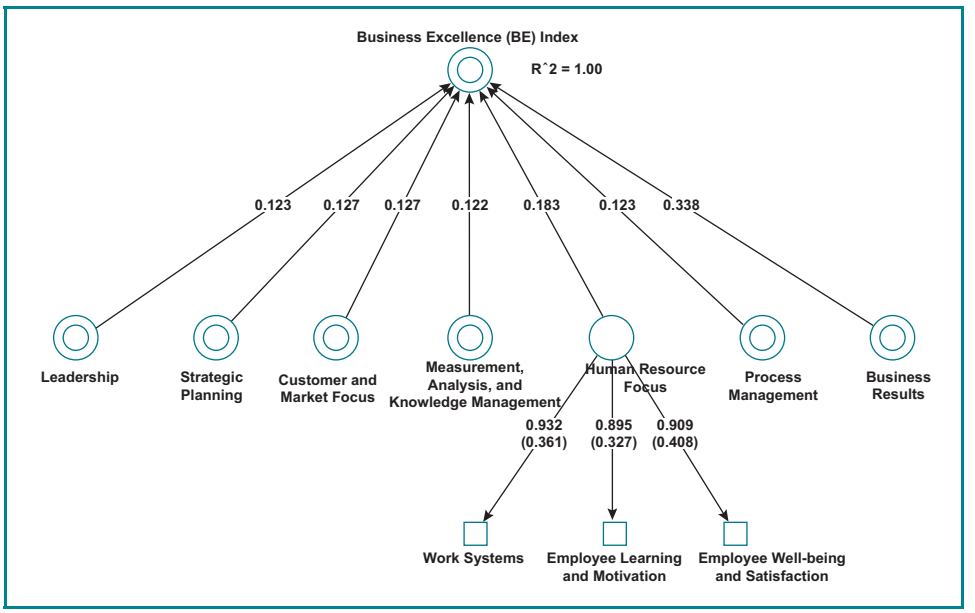
The aforementioned studies on the validity of BE models relate to studies that directly validate the measures (in psychometrics, construct validity) of a BE model, and such studies are our main focus. However, the validity of BE models can be indirectly assessed by testing the association between the BE score (or a similar predictor such as quality award winners) and an external criterion (in psychometrics, criterion-related validity) such as company growth. One of the most well-known indirect tests on the validity of BE models was reported by Hendricks and Singhal (1997); they showed that quality award winners significantly outperform others (the reference group) on a variety of operating performance measures (e.g. operating income, sales growth). Other similar studies include those of Easton and Jarrell (1998), Jacob *et al.* (2004), and Ramasesh (1998).

Given that the BE models are evolving continuously – recently there has been a major emphasis on application of the BCPE in the non-profit sector (Miguel, 2008) – in terms of their operational definitions, researchers need new evidence on the validity of the measures used in BE models. Our literature review shows very little new evidence on the validity of the measures used in BE models (direct validity or indirect validity, based on methods accepted in quantitative research). Our study attempts to address this knowledge gap by using conceivably the most authentic operational definition of BE – being assessments made by trained assessors under national quality award settings.

3. Research questions

Based on the prior literature it has been shown that the categories of a BE model can be modelled either as constructs that predict and explain results (Figure 2), or as constructs (or indicators) of the overall organisational concept Business Excellence (Figure 3). First, the existence of a valid theoretical model that explains results gives the academic community some confidence that there is an underlying theory (a theoretical model may also give confidence to practitioners and policy makers that a BE model can be used for intervention purposes). Second, the existence of a theoretical model that shows all the categories of a BE model form an underlying concept called BE and that the summation of item and category scores provides a valid BE score, gives further confidence to academics and practitioners. Thus we believe that a BE model should fulfil both requirements simultaneously. To the best of our knowledge, empirical research that examines both these requirements simultaneously has not been conducted before. Our aim is to fill this void, using data sets pertaining to

Figure 3 The standardised structural regression coefficients along with item weights (shown in parentheses) of a selected construct (Model II)



applicants for national quality/BE awards in Australasia. We pose the following two measurement questions:

1. Do the measurement items of the BCPE appear to be measuring the seven categories any differently when the categories are being viewed as constructs of a model that predicts and explains results (Figure 2) than when the categories are being viewed as constructs of the overall organisational concept ''Business Excellence'' (Figure 3)?
2. If the measurement items do not appear to be fitting the underlying categories particularly well in either of the aforesaid conceptualisations, what could be the possible reason/s for this?

4. Methodology

In order to answer the first research question, we collected the scores given by each assessor ($n = 118$) for each measurement item for each applicant who applied for the New Zealand Business Excellence Award (NZBEA) during the 2003-2006 period. The NZBEA is based on the Baldrige CPE, and during this period there were 22 applicants; 80 per cent belonged to the service sector, of which two thirds were non-profit organisations. The data were provided by the New Zealand Business Excellence Foundation (NZBEF), a non-profit organisation that administers the NZBEA. However, richer context-bound information such as what the applicants actually submitted alongside their application as evidence of what they do and achieve and assessor reports, which are useful to answer the second research question, were not supplied to us by the NZBEF for confidentiality reasons. However, such information was supplied by SAI Global Australia, who administer the Australian Business Excellence Award (ABEA). Although the ABEA is based on the Australian Business Excellence Framework (ABEF), this framework is quite similar to the BCPE (Saunders and Mann, 2005). In our analysis we assumed that assessments made by the *assessors* of the applicants for the NZBEA are more objective, compared to other forms of assessments such as self-assessment. The 2006 version of the BCPE has been taken as the reference model because the BCPE during 2003-2006 accurately reflected the 2006 version. Hence the BCPE item and category names used herein have been labelled in accordance with the 2006 version.

Both theoretical models that we used (Figures 2 and 3) were tested using the PLSBSEM technique using the PLS-Graph 3.0 software package (Chin, 2001). The results for the first theoretical model (Figure 2) have been reported in Jayamaha *et al.* (2009). The type of theoretical model depicted in Figure 3 is known as a ''second-order molar model'' (Chin and Gopal, 1995); in second-order molar models the second order construct – in our case labelled ''BE'' – is formed by its constituent first-order constructs, much the same way as a mole is formed by its constituents in chemistry (Chin and Gopal, 1995; Dibbern, 2004). In predictive terms, one can view this situation as a second-order response variable being predicted by its first-order predictor variables. The computational algorithm was set such that the total variance of the second-order construct was fully explained ($R^2 = 1$) by the seven first-order constructs, in order to take into account the fact that the overall BE score of an organisation is treated as the sum of the category scores in performance measurement.

The PLBSEM algorithm generates factor scores of the constructs. These scores are used to compute the loadings and cross-loadings. The loading is the correlation between a measurement item and its assigned construct, while a cross-loading is the correlation between a measurement item and any other construct to which the measurement does not belong (Barclay *et al.*, 1995; Chin, 1998; Gefen and Straub, 2005).

5. Results and discussion

Tables I and II depict the loadings (the highlighted correlation coefficient in each row) and cross-loadings (the non-highlighted six correlation coefficients in each row) corresponding to the measurement items of Models I and II, respectively. It is clearly evident from the

Table I Loading and cross-loading pattern of the measurement items for Model I (the model that predicts/explains results)

	Construct no.								
Item no.	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*Average cross-loading*	*ΔV*
1.1	**0.93**	0.78	0.54	0.79	0.73	0.70	0.66	0.70	0.23
1.2	**0.91**	0.67	0.63	0.68	0.76	0.53	0.67	0.66	0.26
2.1	0.76	**0.94**	0.48	0.79	0.72	0.59	0.69	0.67	0.27
2.2	0.74	**0.95**	0.57	0.83	0.78	0.79	0.79	0.75	0.20
3.1	0.67	0.58	**0.95**	0.52	0.58	0.48	0.62	0.58	0.38
3.2	0.53	0.48	**0.95**	0.47	0.50	0.54	0.66	0.53	0.42
4.1	0.76	0.81	0.52	**0.93**	0.82	0.71	0.75	0.73	0.20
4.2	0.67	0.71	0.42	**0.89**	0.66	0.61	0.53	0.60	0.29
5.1	0.79	0.73	0.50	0.79	**0.93**	0.55	0.54	0.65	0.28
5.2	0.67	0.63	0.51	0.71	**0.89**	0.41	0.50	0.57	0.32
5.3	0.74	0.76	0.56	0.77	**0.92**	0.78	0.74	0.72	0.20
6.1	0.54	0.64	0.33	0.64	0.53	**0.91**	0.74	0.57	0.34
6.2	0.69	0.70	0.64	0.70	0.69	**0.93**	0.78	0.70	0.23
7.1	0.53	0.60	0.70	0.51	0.44	0.60	**0.74**	0.56	0.17
7.2	0.66	0.76	0.58	0.65	0.56	0.79	**0.90**	0.67	0.24
7.3	0.52	0.59	0.55	0.60	0.51	0.69	**0.86**	0.57	0.29
7.4	0.66	0.69	0.51	0.60	0.68	0.72	**0.87**	0.64	0.23
7.5	0.61	0.74	0.55	0.69	0.61	0.79	**0.91**	0.66	0.24
7.6	0.67	0.59	0.57	0.57	0.56	0.60	**0.78**	0.59	0.19
Average ΔV									0.26

Note: All correlations (i.e. loadings and cross-loadings) are significant at $\alpha = 0.05$. ΔV is an arbitrary variable that shows by how much a loading exceeds the average cross-loading. The names of the categories and items are as follows: Category 1, Leadership; Category 2, Strategic Planning; Category 3, Customer and Market Focus; Category 4, Measurement, Analysis and Knowledge Management; Category 5, Human Resource Focus; Category 6, Process Management; Category 7, Business Results; Item 1.1, Senior leadership; Item 1.2, Governance and social responsibilities; Item 2.1, Strategy development; Item 2.2, Strategy deployment; Item 3.1, Customer and market knowledge; Item 3.2, Customer relationships and satisfaction; Item 4.1, Measurement, analysis, and review of organisational performance; Item 4.2, Information and knowledge management; Item 5.1, Work systems; Item 5.2, Employee learning and motivation; Item 5.3, Employee wellbeing and satisfaction; Item 6.1, Value creation processes; Item 6.2, Support processes and Operational planning; Item 7.1, Product and service outcomes; Item 7.2, Customer-focused results; Item 7.3, Financial and market results; Item 7.4, Human resource results; Item 7.5, Organisational effectiveness results; Item 7.6, Leadership and social responsibility results
Source: Jayamaha *et al.* (2009)

Table II Loading and cross-loading pattern of the measurement items for Model II (the model that conceptualises business excellence as second-order construct)

Item no.	*Construct no. 1*	*2*	*3*	*4*	*5*	*6*	*7*	*Average cross-loading*	*ΔV*
1.1	**0.92**	0.78	0.54	0.79	0.73	0.70	0.66	0.70	0.22
1.2	**0.92**	0.67	0.63	0.68	0.77	0.53	0.67	0.66	0.26
2.1	0.76	**0.94**	0.48	0.79	0.71	0.59	0.69	0.67	0.27
2.2	0.73	**0.95**	0.57	0.83	0.78	0.79	0.79	0.75	0.20
3.1	0.67	0.58	**0.95**	0.52	0.58	0.49	0.63	0.58	0.37
3.2	0.54	0.48	**0.95**	0.47	0.50	0.54	0.66	0.53	0.42
4.1	0.78	0.84	0.52	**0.93**	0.86	0.71	0.75	0.74	0.19
4.2	0.67	0.71	0.42	**0.89**	0.65	0.61	0.53	0.60	0.29
5.1	0.80	0.73	0.49	0.79	**0.93**	0.55	0.54	0.65	0.28
5.2	0.68	0.63	0.51	0.71	**0.90**	0.41	0.50	0.57	0.32
5.3	0.74	0.78	0.56	0.79	**0.91**	0.79	0.74	0.73	0.18
6.1	0.53	0.64	0.33	0.64	0.52	**0.91**	0.74	0.57	0.34
6.2	0.69	0.70	0.64	0.70	0.68	**0.93**	0.79	0.70	0.23
7.1	0.54	0.60	0.70	0.51	0.44	0.60	**0.74**	0.56	0.18
7.2	0.66	0.76	0.58	0.65	0.55	0.79	**0.90**	0.67	0.23
7.3	0.52	0.59	0.55	0.60	0.51	0.69	**0.86**	0.57	0.29
7.4	0.66	0.69	0.51	0.61	0.67	0.72	**0.87**	0.64	0.22
7.5	0.61	0.74	0.55	0.69	0.60	0.79	**0.90**	0.66	0.24
7.6	0.67	0.59	0.57	0.57	0.56	0.60	**0.78**	0.59	0.19
Average *ΔV*									0.26

Note: All correlations (i.e. loadings and cross-loadings) are significant at $\alpha = 0.05$

correlations depicted in the tables that the loading and cross-loading pattern of the measurement items is nearly the same for both models. Although the loadings are quite strong ($r > 0.7$), suggesting that the measurement items may belong to its assigned category, thus resulting in high convergent validity (Gefen and Straub, 2005), the cross-loadings are also nearly as high as the loadings, suggesting that a measurement item which was thought to belong to a given category (on the strength of high loadings) may in fact belong to the other six categories also, thus demonstrating low discriminant validity (Gefen and Straub, 2005). For example, as shown in Table I, item 1.1 is strongly correlated with Category 1 (the assigned category of item 1.1) with a correlation coefficient of 0.93 (this is item 1.1's loading), but at the same time this measurement item is also highly correlated with some of the other six categories with correlation coefficients of 0.78 (with category 2), 0.54 (with category 3), 0.79 (with category 4), 0.73 (with category 5), 0.70 (with category 6) and 0.66 (with category 7) (these are item 1.1's cross-loadings).

The ΔV criterion used in Tables I and II and defined by us in Jayamaha *et al.* (2009) is a summary statistic that shows how low the cross-loadings are compared to the loading; as a basic requirement for measurement validity we recommended that ΔV should be at least 0.20, but we also recommended other heuristics on cross-loadings to test the level of measurement validity more closely; these heuristics require most of the cross-loadings to be less than 0.60 for a satisfactorily high level of measurement validity (a high cross-loading fails to discriminate/differentiate a measurement item between its assigned category and other categories), which obviously neither of the two models (Model I and Model II) satisfies. Thus, based on our heuristics (Jayamaha *et al.*, 2009) and other rules of thumb used in discriminant validity testing (Barclay *et al.*, 1995; Chin, 1998; Gefen and Straub, 2005), both models show equally low levels of measurement validity.

As stated earlier, the measurement validity of both models is important. Model I legitimises the existence of a theory on organisational performance excellence underlying the BCPE. Model II legitimises adding the scores of measurement items to compute the overall BE score for recognition (e.g. national awards), benchmarking and sharing of best practices. Model II can also be used to test the fairness of the weights stipulated for items and categories of the BCPE in the setting in which it is being used, with the aid of standardised

structural regression coefficients associated with the seven constructs and the items' weights estimated by the PLSBSEM (Figure 3). Our analysis shows that the measurement items do not fit into either of the two conceptualisations particularly well, thus devaluing the usefulness of these models for understanding BE, which brings us back to the second research question. Since we tested the validity of measurement items directly, based on assessments made by trained assessors, we assumed that the reason for low validity of measurement items may be found by examining the way in which organisations have responded to the stipulated requirements on each item.

We were able to gain insights as to how evidence on achievements (approach, deployment etc.) for each measurement item is actually being assessed based on the evidence furnished by applicants for the ABEA (like applicants for the NSBEA, the majority of the applicants for the ABEA were non-profit organisations). In this paper we summarise our findings using measurement item 6.1 (value creation) as a reference item. This measurement item reflects ''primary activities'' or value-bearing activities referred to in Michel Porter's ''value chain'' model (Porter, 1998, p. 37). The reader may note that measurement item 6.1 shows a higher level of measurement validity compared to most of the other measurement items (those whose ΔVs are closer to, or less than 0.20).

Table III summarises responses made by three ''Silver Award Winners'' of the ABEA, in response to requirements stipulated under the measurement item ''innovation process'' in the ABEF (SAI Global, 2004). This item is similar to item 6.1 in the BCPE. We selected the three award winners (all were non-profit organisations) because the evidence supplied by these high scoring organisations can be assumed to be true and accurate, based on the rigorous evaluation process they had been through and the additional checks that would have been undertaken for potential award winners. From Table III, it is evident that although the measurement item under observation is required to belong to the category Process Management, it might also belong to several other constructs such as:

- Measurement, Analysis, and Knowledge Management;
- Human Resource (Workforce) Focus; and
- Leadership.

We observe that most of the measurement items show a similar trend (i.e. appearing to belong to several constructs), based on the supporting evidence the three organisations have furnished. We found this pattern to also be true in other application documents from non-award winners. We note that during the assessment process, examiners look for alignment between the categories and evidence of an integrated approach that cuts across all categories, and hence it is not surprising that organisations produce evidence of approach, deployment, learning and integration of processes (under each item) that appear to belong to multiple categories (Blazey, 2002; National Institute of Standards and Technology, 2009). We thus propose that the main reason for low discriminant validity of measures of the BCPE is attributable to criteria requirements more than anything else.

Another interesting observation we make is that while all three organisations have addressed the first requirement (systems are in place to harvest creative ideas) under the measurement item ''innovation process'', the three non-profit organisations have not addressed some of the other requirements particularly well (e.g. as evidenced from Table III, organisation C does not seem to have responded to requirements 3 and 5). Indeed, from exploratory research of the responses in comparison to other responses from other organisations from other sectors it appears that the business sector within which an organisation operates may have an impact on the relevance of an item. However, we were unable to investigate this further due to a shortage of data (we would need to undertake a cross-comparison of organisational responses for multiple items based on large samples comprising of both for-profit and non-profit organisations and this was beyond the scope of our paper). There is literature that indicates that the business excellence models should be different for different business sectors. For instance, it has recently been acknowledged that performance measurement of a non-profit organisation is far more complex than a ''comparable-sise'' for-profit organisation (Foster *et al.*, 2009; van Iwaarden *et al.*, 2009).

Table III Organisational responses to the requirements stipulated for the item ''innovation process'' of the ABEF

Stipulated requirements/evidence	*Organisation A*	*Organisation B*	*Organisation C*
1. Evidence to support that mechanisms are in place to harvest creative ideas and that top management truly values creative ideas – which may come from within (e.g. employees) or outside (e.g. clients/customers)	Encourage employees (at all levels) to contribute ideas The IT system is geared to track high leverage customer service opportunities based on customer complaints/suggestions (the decision support system factors such indicators as type of customer, nature/impact of customer suggestion, suggestion frequency etc.) Incentive schemes put in place to induce employees to contribute with high leverage customer service or process management opportunities Establishment of strategic partnerships with external bodies Employees who have greater interface with customers are issued with ''quality assurance certificates'' Maintains high level of customer (community) consultation The CEO provides immense support and encouragement for employee-initiated innovation	Existence of an integrated information management system Staff having free access to the organisation's knowledge repository A six-member innovation team to handle low leverage ideas; productivity improvement teams (PITs) to handle high leverage ideas The ''employee-idea generation'' is referred to by a metaphor to energise organisational members; leadership values employee ideas The organisation promotes autonomy; low leverage ideas are rarely being referred to top management; almost invariably, these ideas are implemented by ''third line'' management Leadership share achievements (resulting from suggestions) with employees in publicising success stories The organisation is overwhelmed by the number of volunteers wanting to join PITs	Existence of an integrated information management system that tracks the progress of suggestions Existence of an intranet system with access to all staff members in order to improve process efficiency Appointment of innovation mentors Committees are put in place to direct towards best practice
2. Evidence to suggest that adequate resources are allocated to further enhance idea generation.	Enhance the existing knowledge base through strategic HR activities (recruitment, organisation-specific training etc.)	Leaderships assign a facilitator and a team leader for each PIT to promote learning and continuous improvement Resources for innovation are integrated into the strategic plan	Extensive training of staff in order to harvest a greater number of ideas Resources for innovation are integrated into the corporate plan
3. Evidence to support that ideas are screened by the management to identify high leverage opportunities	The decision support system described above The CEO is also the patron for the community consultation group	Leadership scrutinise the ideas suggested by the PITs	No response
4. Evidence to support that R&D activities are undertaken to explore new products/services/technologies, etc.	Technological developments are monitored all the time Staff are encouraged to exchange knowledge with their counterparts (socialising) in seminars relevant to the industry	Activities of the PITs are implicitly treated as R&D activity; provides statistics on PITs' performance (e.g. number of ideas generated, number of ideas actually implemented, no. of people actively participating in PITs, number of facilitators, number of PITs appointed per annum, etc.)	Extensive collaboration with universities (social science scholars) to learn more about institutionalisation and how the organisation can be improved to meet customer/community expectations
5. Evidence to support process innovation and/or optimisation methods designed to meet customer expectations	''First of its kind in Australia'' type civil engineering technology application described. Energy use optimisation methods described	The ''minutes of meetings'' of the PITs are furnished for scrutiny	No response
6. Evidence of product/service innovation; the process of gathering and analysing market intelligence such as predicting customer expectations, may also be included by the applicants	Engineering and architectural designs are furnished for scrutiny	The ''minutes of meetings'' of the PITs are furnished for scrutiny	A sophisticated forecasting model is put in place to track and forecast changes in the business environment (e.g. demographics, community outcomes that impact the operations of the organisation)

In summary, our empirical analysis on the analysis of correlations suggested that measurement items do not fit into their assigned categories adequately, in the sense that most of the items would have fitted to categories other than their assigned categories (high cross-loadings). We suggested that the main reason for low discriminant validity of measures of the BCPE is attributable to criteria requirements in having to have evidence of an integrated approach that cuts across all categories.

6. Conclusions

Two important measurement perspectives of a BE model were tested empirically by using the BCPE as a reference model, with data from national BE awards (quantitative data from New Zealand and qualitative data from Australia). In national quality award settings, organisations commit vast amounts of resources to furnish evidence on what they do and what they achieve, in response to each assessment area in a BE model. It is widely believed that performance measurement under such settings is quite useful to organisations for:

1. intervention purposes to understand how results are caused; and
2. benchmarking, sharing of best practices and recognition (Grigg and Mann, 2008).

One way to further enhance one's understanding about how results are caused is to have recourse to the deductive approach of testing a theoretical model that represents how results are caused (Figure 2). In the same vein, one can test whether or not adding items scores to constitute an overall BE score – which is used for benchmarking, sharing of best practices and recognition – makes empirical sense, using an appropriate theoretical model (Figure 3). Essentially these are two different measurement perspectives and we tested both of them using the same data. We found that most of the measurement items of the BCPE do not fit either of the two measurement perspectives particularly well (this answered our first research question) and we proposed that the main reason for this is attributable to criteria requirements in having to have evidence of an integrated approach that cuts across all categories (this answered our second research question).

Developing discrete measures (for high discriminant validity) to operationalise abstract concepts (e.g. leadership) is of paramount importance in quantitative research. However, from a theoretical perspective, BE models – which are practitioner oriented – impose a challenge. This is because in BE models all measurement items have been designed to be related to each other as this is considered to be a feature of business excellence and when an organisation implements management/improvement interventions it is usually desired that they impact on multiple items. From a practical perspective, working with measurement items that are discrete, if they could be so designed, could have a negative consequence for business excellence. For example, organisations may become obsessed with focussing on implementing interventions that improve a particular item's score, and yet these may not have an impact on other items and therefore their impact on overall business excellence may be limited.

The opportunity for researchers is to learn more about the inter-relationships between the various items and categories. This will assist practitioners who use the models. Users need to understand the holistic nature of the models, the linkages between the various categories and items, and the scoring system. It is also recommended that further research is undertaken to determine whether the design of business excellence models should be different for different industry sectors. Therefore, can one design fit all or should it be tailored? There is a need to conduct an exhaustive literature review on management methods for different sectors as well as cross-comparison of organisational responses (in the case of assessments being made by trained assessors) for multiple items based on large samples comprising of different industry sectors.

References

Anderson, J.C., Rungtusanatham, M. and Schroeder, R.G. (1994), ''A theory of quality management underlying the Deming management method'', *Academy of Management Review*, Vol. 19 No. 3, pp. 472-509.

Barclay, D., Thompson, R. and Higgins, C. (1995), ''The partial least squares (PLS) approach to causal modeling: personal computer adoption and use as an illustration'', *Technology Studies*, Vol. 2 No. 2, pp. 285-309.

Blazey, M.L. (2002), *Insights to Performance Excellence: An Inside Look at the 2002 Baldrige Award Criteria*, ASQ Press, Milwaukee, WI.

Bou-Llusar, J.C., Escrig-Tena, A.B., Roca-Puig, V. and Beltrán-Martín, I. (2005), ''To what extent do enablers explain results in the EFQM excellence model? An empirical study'', *International Journal of Quality & Reliability Management*, Vol. 22 No. 4, pp. 337-53.

Chin, W.W. (1998), ''The partial least squares approach to structural equation modelling'', in Marcoulides, G.A. (Ed.), *Modern Methods for Business Research*, Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 295-336.

Chin, W.W. (2001), *PLS-Graph User's Guide, Version 3.0*, C.T. Bauer College of Business, University of Houston, Houston, TX.

Chin, W.W. and Gopal, A. (1995), ''Adoption intention in GSS: relative importance of beliefs'', *Data Base Advances*, Vol. 26 No. 2, pp. 42-64.

Collier, D.A., Goldstein, S.M. and Wilson, D.D. (2002), ''A thing of the past?'', *Quality Progress*, Vol. 35 No. 10, pp. 97-104.

Curkovic, S., Melnyk, S., Calantone, R. and Handfield, R. (2000), ''Validating the Malcolm Baldrige National Quality Award framework through structural equation modelling'', *International Journal of Production Research*, Vol. 38 No. 4, pp. 765-91.

Dahlgaard, J.J., Kristensen, K. and Kanji, G.K. (1998), *Fundamentals of Total Quality Management: Process Analysis and Improvement*, Chapman & Hall, London.

Dean, M. and Tomovic, C. (2004), ''Does Baldrige make a case for quality?'', *Quality Progress*, Vol. 37 No. 4, pp. 40-5.

Deming, W.E. (1986), *Out of the Crisis*, MIT Press, Cambridge, MA.

Dibbern, J. (2004), *The Sourcing of Application Software Services: Empirical Evidence of Cultural, Industry and Functional Differences*, Physica-Verlag, Heidelberg.

Easton, G.S. and Jarrell, S.L. (1998), ''The effects of total quality management on corporate performance: an empirical investigation'', *The Journal of Business*, Vol. 71 No. 2, pp. 253-307.

Eskildsen, J.K., Kristensen, K. and Juhl, H.J. (2001), ''The criterion weights of the EFQM excellence model'', *International Journal of Quality & Reliability Management*, Vol. 18 No. 8, pp. 783-95.

Eskildsen, J.K., Kristensen, K. and Juhl, H.J. (2002), ''Trends in EFQM criterion weights: the case of Denmark 1998-2001'', *Measuring Business Excellence*, Vol. 6 No. 2, pp. 22-8.

Flynn, B.B. and Saladin, B. (2001), ''Further evidence on the validity of the theoretical models underlying the Baldrige criteria'', *Journal of Operations Management*, Vol. 19 No. 6, pp. 617-52.

Foster, W.L., Kim, P. and Christiansen, B. (2009), ''Ten nonprofit funding models'', *Stanford Social Innovation Review*, Spring, pp. 32-9.

Garvin, D.A. (1991), ''How the Baldrige Award really works'', *Harvard Business Review*, Vol. 69 No. 6, pp. 80-93.

Gefen, D. and Straub, D. (2005), ''A practical guide to factorial validity using PLS-Graph: tutorial and annotated example'', *Communications of the Association for Information Systems*, Vol. 16, pp. 91-109.

Grigg, N. and Mann, R. (2008), ''Rewarding excellence: an international study into business excellence award processes'', *Quality Management Journal*, Vol. 15 No. 3, pp. 26-40.

Hendricks, K.B. and Singhal, V.R. (1997), ''Does implementing an effective TQM program actually improve operating performance? Empirical evidence from firms that have won quality awards'', *Management Science*, Vol. 43 No. 9, pp. 1258-74.

Jacob, R., Madu, C.N. and Tang, C. (2004), ''An empirical assessment of the financial performance of Malcolm Baldrige Award winners'', *International Journal of Quality & Reliability Management*, Vol. 21 No. 8, pp. 897-914.

Jayamaha, N.P., Grigg, N.P. and Mann, R.S. (2008), ''Empirical validity of Baldrige Criteria: New Zealand evidence'', *International Journal of Quality & Reliability Management*, Vol. 25 No. 5, pp. 477-93.

Jayamaha, N., Grigg, N. and Mann, R. (2009), ''A study of the validity of three major business excellence models in the Asia Pacific region'', *Total Quality Management & Business Excellence*, Vol. 20 No. 11, pp. 1213-27.

Lee, S.M., Rho, B.H. and Lee, S.G. (2003), ''Impact of Malcolm Baldrige National Quality Award criteria on organisational quality performance'', *International Journal of Production Research*, Vol. 41 No. 9, pp. 2003-20.

Meyer, S.M. and Collier, D.A. (2001), ''An empirical test of the causal relationships in the Baldrige health care pilot criteria'', *Journal of Operations Management*, Vol. 19 No. 4, pp. 403-26.

Miguel, P.A.C. (2008), ''Business excellence through a world-class organisational structure: experience from the Baldrige national quality program'', *International Journal of Business Excellence*, Vol. 1 Nos 1/2, pp. 175-92.

National Institute of Standards and Technology (2009), *Criteria for Performance Excellence 2009-2010: Baldrige National Quality Program*, US Department of Commerce, Gaithersburg, MD.

Pannirselvam, G.P. and Ferguson, L.A. (2001), ''A study of the relationships between the Baldrige categories'', *International Journal of Quality & Reliability Management*, Vol. 18 No. 1, pp. 14-37.

Pannirselvam, G.P., Siferd, S.P. and Ruch, W.A. (1998), ''Validation of the Arizona Governor's Quality Award criteria: a test of the Baldrige criteria'', *Journal of Operations Management*, Vol. 16, pp. 529-50.

Porter, L.J. and Tanner, S.J. (2004), *ECBS Assessing Business Excellence: A Guide to Business Excellence and Self-assessment*, Elsevier, Boston, MA.

Porter, M.E. (1998), *Competitive Advantage: Creating and Sustaining Superior Performance*, The Free Press, New York, NY.

Prajogo, D.I. and Brown, A. (2004), ''The relationship between TQM practices and quality performance and the role of formal TQM programs: an Australian empirical study'', *QMJ*, Vol. 11 No. 4, pp. 31-42.

Ramasesh, R.V. (1998), ''Baldrige Award announcement and shareholder wealth'', *International Journal of Quality Science*, Vol. 3 No. 3, pp. 114-25.

SAI Global (2004), *Australian Business Excellence Framework*, GB 002 – 2004 ed., SAI Global, Sydney.

Saunders, M. and Mann, R. (2005), ''Self-assessment in a multi-organisational network'', *International Journal of Quality & Reliability Management*, Vol. 22 No. 6, pp. 554-71.

Tamimi, N. (1998), ''A second-order factor analysis of critical TQM factors'', *International Journal of Quality Science*, Vol. 3 No. 1, pp. 71-9.

van Iwaarden, J., van der Wiele, T., Williams, R. and Moxham, C. (2009), ''Charities: how important is performance to donors?'', *International Journal of Quality & Reliability Management*, Vol. 26 No. 1, pp. 5-22.

Williams, R., Bertsch, B., van der Wiele, A., van Iwaarden, J. and Dale, B. (2006), ''Self-assessment against business excellence models: a critique and perspective'', *Total Quality Management & Business Excellence*, Vol. 17 No. 10, pp. 1287-300.

Further reading

Chin, W.W. (1998), ''Commentary: issues and opinion on structural equation modelling'', *MIS Quarterly*, Vol. 22 No. 1, pp. vii-xvi.

About the authors

Nihal Jayamaha is a Lecturer in Quality Management at Massey University. He previously worked in large electrical utilities in South Asia and the Middle East for 20 years as a Project Manager, an Operations Engineer and an Auditor. He has published several journal articles, book chapters and conference papers on TQM and business excellence. His research interests include industrial process modelling, performance measurement, quality management in healthcare and TQM theory building for the non-profit sector. He obtained his PhD in Technology from Massey University in 2009. He also holds a Masterate in Engineering (ME) and a Masterate in Business Administration (MBA). Nihal Jayamaha is the corresponding author and can be contacted at: n.p.jayamaha@massey.ac.nz

Nigel Grigg is an Associate Professor in Quality Systems at Massey University, and is the Coordinator of Massey University's three postgraduate level programmes in quality systems. He is a Chartered Quality Professional and a member of the Chartered Quality Institute, the American Society for Quality and the Institute of Directors. He is currently President and Chairman of the New Zealand Organisation for Quality (NZOQ). His research interests include process improvement and the use of statistical thinking and methods within the knowledge development cycle. He has authored/co-authored over 80 journal and conference papers on various aspects of quality management, and has received three national best journal paper awards, in addition to the 2004 annual National Award for best thesis from the Chartered Quality Institute.

Robin Mann is the Head of the Centre for Organisational Excellence Research, New Zealand (see: www.coer.org.nz), Chairman of the Global Benchmarking Network (see: www.globalbenchmarking.org), Advisory Board member at the Hamden Bin Mohammed e-University, Dubai (see www.hbmeu.ae) and Co-Founder of BPIR.com Limited (see: www.bpir.com), a leading benchmarking website resource. His experience includes managing the UK's Food and Drinks Industry Benchmarking and Self-assessment Initiative (1995-1998), New Zealand Benchmarking Club (2000-2004), the Sheikh SAQR Government Excellence Program, UAE (2005-2007), reviewing the Australian Business Excellence Framework (2006), delivering TRADE best practice benchmarking training and certification with COER's partners in 11 countries, and chief expert for the APO's research project on business excellence in Asia.