Jayamaha, N.P, Mann, R.S. and Grigg, N, 2008. An empirical study of the validity of business excellence models and the relationships between 'enablers' and 'business results. Chapter 1, Quality Management and Organisation Excellence: Oxymorons, Empty Boxes, or Significant Contributions to Management Thought and Practice. Editors Foley, K, Hensler, D.A. and J. Jonker. *Published by: Standards Australia International (SAI) Global.*.

An empirical study of the validity of business excellence models and the relationships between 'enablers' and 'business results'

Nihal P. Jayamaha, Robin S. Mann, Nigel P. Grigg, Massey University, New Zealand.

Abstract

The key purpose of this study was to examine the empirical validity of three key Business Excellence (BE) models used in the Australasia/Asia Pacific region: the Australian Business Excellence Framework (ABEF); the Baldrige Criteria for Performance Excellence (BCPE, used in New Zealand); and the Singapore Quality Award Criteria (SQAC). In addition, the study also assessed the empirical relationships between the *enabler items* (i.e. measures on leadership and performance management) and the *business results items* (i.e. those measures that cover key stakeholder outcomes), using predictive linear models.

The study is original, in the following respects: (a) it is based on data (item scores) on past Australian/New Zealand/Singaporean BE/Quality Award applicants (N = 113 for the SQAC, N = 110 for the ABEF, and N = 118 for the BCPE) and reflects the true *measurement validity*of the BE models; (b) it introduces a generic methodology for evaluating the measurement validity of a BE model; and (c) it introduces the *partial least squares regression* (PLSR) *method* to quality management, through studying the relationships between the enablers (predictor variables) and the results (response variables). The PLSR was used to handle the problem of multicollinearity of the predictor variables (the *ordinary least squares regression method* yielded unstable/un-interpretable regression parameter estimates).The measurement validity of the three models was assessed using the *partial least squares approach to path modelling,* by studying the patterns of loadings and crossloadings between the BE measurement items and the theoretical constructs; this gives evidence (or the lack of it) of the *convergent validity* and the *discriminant validity*.

The three BE models were ranked for level of measurement validity, based on the following three heuristics: (i) the proportion of measurement items with loadings that are at least 0.20 greater than their average crossloadings; (ii) the proportion of cross loadings < 0.60; (iii) the proportion of measurement items with at least, all but one of their crossloadings < 0.60. For the ABEF in particular, it was observed that although the measurement models showed high loadings (hence convergent validity), they also showed reasonably high crossloadings (hence lower than desirable descriminant validity), which implied that there is scope for improvement of the operationalisation of the quality/BE concepts. Specifically the following measurement gaps were identified: the need to improve several measurement items of the ABEF, and the need to improve one measurement item of the SQAC (note that there was no measurement item in the BCPE that called for improvement). PLSR models on the SQAC and the BCPE showed that (based on the magnitude of the standardised regression coefficients) although certain enabler items seem to be more influential than others in predicting results for a certain group of stakeholders (i.e. certain business results items), when results for all the key stakeholders are taken as a whole, all the enabler items become equally influential in predicting the overall organisational results.

Introduction

The aim of the study reported in this chapter was to empirically examine the validity of three key Business Excellence (BE) models used in the Australasia/Asia Pacific region: the Australian Business Excellence Framework (ABEF), the Baldrige Criteria for Performance Excellence (BCPE, used in New Zealand), and the Singapore Quality Award Criteria (SQAC). Pursuant to this general aim, a related objective was to measure the empirical relationships between the *enabler items* (i.e., measures on leadership and performance management) and the *business results items* (i.e., those measures that cover key stakeholder outcomes), using predictive linear models. The study is original in the following respects: (a) it is based on genuine business excellence scoring data (item scores) from past Australian, New Zealand and Singaporean BE/Quality Award applicants (N = 113 for the SQAC, N = 110 for the ABEF, and N = 118 for the BCPE), and thus reflects the true *measurement validity* of the BE models; (b) it introduces a new generic methodology developed from the study for evaluating the measurement validity of a BE model with respect to other models; and (c) it introduces the *partial least squares regression* (PLSR) *method* as an application for quality management research, through studying the relationships between the enablers (predictor variables) and the results (response variables).

Business Excellence (BE) models and the importance of determining their validity

A BE model can be viewed as: (1) an instrument that measures the level of performance management within organization and results achieved for the key stakeholders; and (2) a representation of theory on organization-wide performance improvement (Flynn and Saladin, 2001; Kanji, 2002). While there is no unified theory on BE, it has been established that BE can be "reliably distinguished" from other organization-wide performance management interventions such as participative management and Management By Objectives (Hackman and Wageman, 1995). What sets apart total quality management (TQM) and BE models from other performance management models is that TQM/BE

models are designed to address the whole management philosophy of an organization and the activities it uses to pursue it. TQM/BE models aim to guide organizations to consistently exceed the current and future expectations of all stakeholders (i.e., customers, employees, shareholders and the community) through "continuous improvement in all processes, goods and services" (Sitkin, Sutcliffe, and Schroeder, 1994). Central to this realization is the creation of a working culture (by the senior leadership) that uses data information and knowledge for every sphere of organizational activity, and evolution of a managerial system that fulfils the intrinsic and extrinsic needs of the organization's employees (Dean and Bowen, 1994; Hackman and Wageman, 1995; Kanji and Wallace, 2000).

When an organization applies for a national BE Award, the key areas of organizational capability are assessed against the model and points are allocated to each measurement item by a panel of trained independent evaluators using a scoring guideline based upon the level of evidence of actual performance. The reader is referred to tables 1-3 for a full list of categories and measurement items pertaining to the three BE models covered in this study. The assessment process used by independent evaluators is exhaustive and typically involves scrutinizing records, meetings with senior managers and their subordinates as well as conducting actual observations on key processes (Grigg and Mann, 2008; SAI Global, 2004). Moreover, irrespective of the level of actual achievement, all applicants receive a feedback report, with an outline of the scoring for the applicant organization, describing areas identified as particular strengths or opportunities for possible improvement. Hence, apart from other indirect benefits such as general *kudos* or demonstrating commitment to BE to key stakeholders, organizations can potentially reap direct benefits through being assessed against a BE model, provided it can be demonstrated that the model is a valid measurement instrument that underpins a theory of organizational performance. One-way to achieve this is to empirically determine the validity of national/regional level BE models. It logically follows that if a national/regional level BE model is invalid, any device that is derived from it (e.g., a self-assessment instrument) must also lack validity. Considering the worldwide interest and popularity of BE (e.g., extensive use of self-assessment instruments,

workshops) it remains paramount that national/regional level BE models be validated using empirical data.

A major obstruction that hinders research studies on the empirical validity of national/regional level BE models in general is the absence of available data, normally due to strict confidentiality rules concerning historical data on national BE award applicants (Garvin, 1991; Pannirselvam, Siferd, and Ruch, 1998). There are only a handful of studies (summarised later) that examined the empirical validity of any national BE model, based on data from past award applicants. Our study is aimed primarily at closing this gap. In our empirical study we cover three models: the Australian Business Excellence Framework (ABEF), the Baldrige Criteria for Performance Excellence (BCPE)[1], and the Singapore Quality Award Criteria (SQAC).

This paper is divided into two parts. Part I summarises prior research and discusses our methodology on validity assessment, the findings and the implications of those findings. Part II compliments Part I by demonstrating how linear predictive models on business results can be developed using a leading edge multivariate statistical technique. We believe that such models, which are virtually nonexistent in BE literature, would be useful in demonstrating the extent to which a given measurement item (or a measurement category) can predict or influence business results. By citing an example we have also shown how these models may be used to pinpoint potential problems on implementation of performance improvement practices.

Part I

Facets of validity

[1] It should be noted that BCPE—the most prestigious quality award in USA—has been assessed using data from applicants for the New Zealand Business Excellence Award.

Two types of important (and interrelated) validity tests are applicable to BE models (or any other model that involves measurement of concepts/constructs), namely measurement validity and statistical conclusion validity. Measurement validity (the most fundamental validity upon which other forms of validity are built) refers to demonstration of the extent to which a measurement instrument measures the concepts or constructs that it purports to measure (Nunnally and Bernstein, 1994). Statistical conclusion validity refers to demonstration of the fact that the hypothesized relationships between the constructs, as represented by a model, do exist statistically (MacKenzie, 2003).

Construct validity—which subsumes many other forms of measurement validity—is considered to be the most important form of measurement validity (Kline, 1998; Nunnally and Bernstein, 1994). A measurement instrument is said to possess construct validity (Cronbach and Meehl, 1955) if it can be established that the measurement items that are assumed to belong within a given construct, appear to be belong to that construct only (Straub, Boudreau, and Gefen, 2004). Obviously, it is not possible to test the construct validity without knowing how the respondents scored in each measurement item belonging to each construct.

Prior Research

Previous research aimed at establishing the measurement validity of any major national/regional BE model is not extensive, for wont of available data—in particular the scores obtained by award applicants on the measurement items. An exception is the research by Pannirselvam *et al*., (1998) who used item scores of applicants who applied for a major US state (Arizona) quality award to establish the measurement validity of the BCPE on the grounds that the state quality award criteria they chose were almost identical to those of the 1994 BCPE. Using the same dataset, Pannirselvam and Ferguson (2001) also established the statistical conclusion validity of the BCPE. These two studies remain valuable pieces of evidence as to the validity of early versions of the BCPE.

It is important to note that the BCPE model was substantially revised in 1997, with major changes to the measurement items, conceptual labels assigned to the seven categories and

the structural model that depicts the relationships between the seven categories (or constructs). It is encouraging to note that the basic framework of the BCPE (i.e., the constructs as well as the implied relationships between the constructs) has remained stable since 1997 in spite of the revisions and refinements to the measurement items (which typically occurs every other year).

Others used less stringent tests on measurement validity. Hausner (1999) established the predictive validity of the ABEF by demonstrating that the total ABEF scores of a sample of past Australian Business Excellence Award applicants ($N = 22$) were highly correlated with the rate of improvement of key performance indicators (KPIs) over a period of 8 years. Predictive validity is a form of measurement validity that demonstrates the ability of a measurement instrument to estimate some criterion that is external to the measurement instrument itself (in the case of the aforesaid research the external criterion being the KPIs). In two independent studies conducted using survey data (obtained using questionnaires framed to reflect the performance requirements sought in the ABEF/BCPE), Rahman (2001) and Samson & Terziovski (1999) respectively and indirectly established the measurement validity of the ABEF and the BCPE. Having established measurement validity they studied the bivariate and multivariate relationships between performance management constructs and operational results. These studies showed that the so-called "soft constructs of quality/BE" such as *leadership* and *human resource focus* were more strongly correlated with operational results than the "hard constructs of quality/BE" such as *process management* and *information and analysis*, thus corroborating the findings of an earlier study by Powell (1995). Powell's proposition that "tacit resources, and not TQM tools and techniques, drive TQM success, and that organizations that acquire them can outperform competitors with or without the accompanying TQM ideology" not has only provided fodder for the skeptics to demean TQM/BE but also prompts the BE researchers to bring more sophisticated quantitative methods to investigate empirical claims.

There have been many studies that assessed the measurement validity and statistical conclusion validity of BCPE using Structural Equation Modeling (SEM) methods, predominantly using the covariance based LISREL software, using survey data. While there are inherent weaknesses in survey data, such as common method bias (Podsakoff,

MacKenzie, Lee and Podsakoff, 2003), these studies are nonetheless invaluable in understanding the theoretical framework underpinning BE models, having constructs similar to the BCPE framework. Studies that attempted to establish the statistical conclusion validity of early versions (pre 1997) of the BCPE framework (e.g., Handfield and Ghosh, 1995; Wilson and Collier, 2000; Winn and Cameron, 1998) reported mixed results ending up with models having several statistically nonsignificant paths, suggesting little or partial evidence of the existence of the causal hypotheses implied in the BCPE framework. However, by switching from confirmatory mode to exploratory mode (using the "theory trimming" technique in LISREL), Winn and Cameron were able to devise an alternative statistically significant model. The alternative model implied that leadership does not have a direct effect on organizational outcomes but only an indirect effect through the systems and processes conceptualized in the other Baldrige constructs. We note that several relationships implied in Winn and Cameron's derived structural model are implied in the post-1997 BCPE framework and that Winn and Cameron's derived structural model has been cross-validated in some subsequent studies (e.g., Badri *et al*., 2006; Meyer and Collier, 2001). In our opinion, however, Winn and Cameron's derived model does not fully reflect the post 1997 BCPE, both in terms of operationalisation of the constructs as well as the implied causal relationships between the constructs.

Using survey data collected from a multinational sample, Flynn and Saladin (2001) studied the goodness-of-fit of three path models—respectively representing the 1988, 1992 and 1997 versions of the BCPE—in order to make a normative assessment of how the BCPE has evolved at a theoretical level. They studied the path models using path analysis, which is a more basic form of SEM. They observed that the 1997 model was a better fit to data (in terms of that model's ability to imply the observed correlations between the measurement items) compared with the 1988 and 1992 models (1992 model was a better fit to data than the 1988 model), which prompted them to conclude that "appropriate modifications have been made to the criteria upon its inception in 1988." Using survey data from a Korean manufacturing sample Lee, Rho, and Lee (2003) established the statistical conclusion validity of the BCPE using SEM, adopting a structural model similar to that used by Flynn and Saladin for the 1997 model.

Methodology

At the outset of this study, we believed that it is more informative and enriching to assess the empirical validity of a multitude of conceptually analogous BE models—that is BE models having constructs with near one-to-one correspondence—within a single theoretical framework. We collected scores (at item level) of past award applicants of the New Zealand Business Excellence Award (like many other countries, the New Zealand award is based on the BCPE), the Australian Business Excellence Award, and the Singapore Quality Award through the award custodians: New Zealand Business Excellence Foundation (Auckland, New Zealand), SAI Global (Sydney, Australia) and Standards, Productivity and Innovation Board of Singapore (SPRING) respectively.

A major motivation for us to study three different models simultaneously was that although there was near one-to-one correspondence between the categories as theoretical concepts or constructs, there was sufficient disparity between the models in the manner in which the categories were measured, both in terms of the measurement items used, and how each item was measured (operationalised). For example, SPRING Singapore adopts a highly structured approach to assess each measurement item using a questionnaire design (the responses furnished by the applicants are nonetheless verified by trained independent examiners through site visits) while SAI Global adopts (for the ABEF) a more open ended design, suggesting what the respondents may consider furnishing as documentary evidence for accomplishment under each measurement item. We believe that our findings would be useful to both practitioners and academia (not to mention the potential benefit to the award custodians) as our research conclusions are derived from *triangulation*—converging evidence based on different datasets, different measurement instruments, different settings and indeed different statistical techniques.

Since few organizations apply for awards in a calendar year, either by design (as in Singapore) or otherwise (as in Australasia), we adopted the following strategies to increase the number of valid observations. In the case of Australia, we pooled data (final consensus scores) across seven years (1999-2006), which resulted in 110 observations. We believe that this pooling has minimal adverse effect because (1) the ABEF remained unchanged

throughout the study period and (2) that only a negligible number of organizations made repeat applications. For the New Zealand case, we pooled individual evaluator scores (on average about 5 evaluators per applicant) across four years (2003-2006), which resulted in 118 observations. In this case pooling was necessary because there were only 22 applicants, and we could not use previous years' applicants as the BCPE measurement items used in those years differed. In the case of Singapore, we used final consensus scores of all the 113 Singaporean organizations that were evaluated by SPRING during 2004-2005, for performance excellence. Note that in a strict sense, these organizations are not the applicants of the Singapore Quality Award (SQA); only organizations that secure an overall score of 700 points (out of a possible 1000) in performance excellence assessments are invited to apply for the SQA (for details of the SQAC see SPRING, 2007). Gaining qualifications to apply for the SQA is deemed as a journey in Singapore, with organizations having to gradually progress toward the 700 mark from where currently they are.

Of our three samples, the Australian and New Zealand samples showed similarity in terms of the sector to which the organizations belonged (>80% service sector, of which the majority were state owned), the size[2] of the organization (approximately 1/3 medium and 2/3 large). The Singaporean sample had a relatively stronger manufacturing sector representation (42%); the size of the organizations in the Singaporean sample was similar to that of the Australasian samples although the private sector was better represented in the Singaporean sample.

We used the above datasets to run three separate partial least squares (PLS) based structural equation models using a common structural model (see Figures 1 - 3). The common structural model used in our PLS analysis was based on the structural models used by Jayamaha, Grigg and Mann (2008), which is consistent with the models used by Flynn and Saladin (2001) as well as by Lee, Rho and Lee (2003) to study the validity of the BCPE. Note that while the structural models of our three PLS path

[2] The size of the organization is based on the fulltime labour employed and is in accordance with the classification used by the Australian Bureau of Statistics: 'Small' if the organization has less than 20 employees, 'Medium' if the organization has between 21 and 200 employees, and 'Large' if the organization has more than 200 employees.

models were identical, overall the three models were different from one another in that each had different measurement models (i.e., the diagram showing the relationships between categories and measurement items). PLS technique is particularly suitable for running SEM models with small samples/and or in situations where the theoretical concepts are at early stages of development (Chin, 1998; Fornell and Cha, 1994).

The objective function of the partial least squares (PLS) algorithm is the minimization of the residual variances specified in the measurement and structural models (using a series of *least squares* regression equations). Although "all the residual variances are minimized jointly", the PLS optimization algorithm remains "partial", in a least squares sense, in that there is no global criterion being set up for optimization as in the LISREL approach (Wold, 1980, p.67). The computation of PLS model parameters (in both the measurement models and the structural model) is accomplished by an iterative algorithm designed to estimate the score of each construct (constructs are often known as 'latent variables' in SEM) using the assumption that each construct can be expressed as a weighted linear combination of its measurement items (also known as 'indicators' or 'manifest variables' in SEM). Thus estimation of item weights is one of the most unique features of the PLS approach. The application of the PLS technique in studying the validity of BE models is available elsewhere (e.g., Cassel, Hackl, and Westlund, 2000; Kanji and Wallace, 2000; Kristensen, and Eskildsen, 2006; Rosa, Saraiva, and Diz, 2003). Our PLS runs were conducted using PLS Graph version 3.0 software package (Chin, 2001).

In PLS, the construct validity (and hence the measurement validity) is assessed through the convergent and discriminant validities (both forms of validities are vital for construct validity). Convergent validity is said to exist when each measurement item correlates strongly with its designated construct (these correlations are referred to as loadings), while discriminant validity is said to exist when measurement items correlate less strongly with the constructs to which they are not designated (these correlations are referred to as cross-loadings, Gefen and Straub, 2005). The correlations pertaining to the three models are shown in Tables 1 - 3. The structural models of our three frameworks are shown in Figures 1 - 3 (note that the significance statistics reported in the figures are based on the non-parametric bootstrap resampling procedure, as no parametric assumptions are made in PLS).

Considering the fact that one of the main objectives of our study was to assess the measurement validity of the BE models, we were mindful of the potential pitfalls on over-reliance on a SEM approach (in our case PLS). In all hybrid SEM models, the measurement model is linked to the structural model and *vice versa,* and if the structural model does not represent the reality (as we posit it would), any inference made in relation to the measurement model would be questionable. We observe that in the case of the BCPE—one of the models that is scrutinized frequently—researchers still tend to have their own interpretation on how the BCPE framework (see NIST, 2006, p.5) could be represented as a recursive structural model. In order to circumvent this problem, as a secondary measure, we computed the loadings and cross-loadings associated with the measurement items using the Principal Components Analysis (PCA) method. Table 6 in Appendix 1 depicts the loadings and cross-loadings associated with the measurement items of the SQAC. In the PCA procedure we made the assumption that a construct can be approximated by the first principal component derived from its indicators (we observed that the Eigenvalues of second and higher principal components were invariably very low). We conducted the PCA using STATISTICA 6.0 software package.

Results and discussion

Measurement validity

Careful examination of the patterns of correlations in Tables 1 - 3 reveals that in the case of all 3 BE models, the measurement items show strong loadings (the loadings are highlighted for ease of reference), thereby showing convergent validity; the possible exception to this is item 5.1 of the SQAC (Table 3). While the measurement items of the three BE models show convergent validity, it is clearly evident that it is difficult to find concrete evidence of discriminant validity, since the measurement items also show strong to moderately strong cross-loadings. If a measurement item returns cross-loadings that are as strong as its loading, then it becomes difficult to justify that the measurement item under observation is a proper operationalization of its assigned construct as the item might well belong to other constructs (Barclay, Thompson, and Higgins, 1995). However, in PLS, there is no

maximum rule-of-thumb cross-loading cut-off value for discriminant validity (some users such as Barkley *et. al*., suggest 0.50 as an ideal maximum value to keep in line with the cut off value often used in exploratory factor analysis). However, it is deemed mandatory that a measurement item should return a strong loading (ideally > 0.71) that is greater than any of its cross-loadings (Chin, 1998; Fornell and Larcker, 1981; Gefen and Straub, 2005). It is clear from the loading and cross-loading data reported (Tables 1 - 3) that the three BE models fulfill this minimum requirement for validity.

Table 1: Loadings and Cross-Loadings for the ABEF Based on Item Scores Secured by the Applicants of the Australian Business Excellence Award

Latent variable/ Category Item	1	2	3	4	5	6	7	Average Cross-loading	**ΔV**
1.1	**0.93**	0.84	0.73	0.80	0.74	0.74	0.76	0.77	0.16
1.2	**0.92**	0.75	0.76	0.82	0.68	0.73	0.73	0.75	0.18
1.3	**0.93**	0.75	0.75	0.83	0.68	0.73	0.73	0.75	0.19
1.4	**0.67**	0.59	0.54	0.51	0.45	0.49	0.53	0.52	0.15
2.1	0.78	**0.90**	0.73	0.68	0.73	0.75	0.72	0.73	0.17
2.2	0.81	**0.92**	0.73	0.76	0.76	0.74	0.70	0.75	0.17
2.3	0.63	**0.80**	0.67	0.60	0.65	0.65	0.57	0.63	0.17
3.1	0.73	0.76	**0.93**	0.65	0.65	0.80	0.73	0.72	0.21
3.2	0.75	0.71	**0.90**	0.64	0.62	0.73	0.74	0.70	0.20
3.3	0.66	0.68	**0.83**	0.66	0.68	0.73	0.65	0.68	0.15
4.1	0.82	0.74	0.73	**0.92**	0.64	0.68	0.73	0.76	0.20
4.2	0.78	0.71	0.69	**0.91**	0.61	0.66	0.64	0.68	0.23
4.3	0.64	0.61	0.49	**0.81**	0.53	0.53	0.55	0.56	0.25
5.1	0.66	0.79	0.63	0.63	**0.91**	0.71	0.63	0.68	0.24
5.2	0.68	0.73	0.67	0.61	**0.95**	0.72	0.58	0.67	0.29
5.3	0.71	0.74	0.72	0.61	**0.91**	0.74	0.60	0.69	0.22
6.1	0.65	0.67	0.72	0.64	0.69	**0.82**	0.71	0.68	0.14
6.2	0.70	0.75	0.69	0.66	0.68	**0.85**	0.65	0.69	0.16
6.3	0.64	0.67	0.73	0.58	0.64	**0.86**	0.65	0.65	0.21
6.4	0.70	0.69	0.78	0.58	0.67	**0.89**	0.74	0.70	0.20
7.1	0.77	0.72	0.75	0.71	0.64	0.76	**0.94**	0.73	0.22
7.2	0.72	0.70	0.74	0.66	0.58	0.74	**0.93**	0.69	0.24
Average ΔV									0.20

Note: (1) ΔV is an arbitrary variable, which shows by how much a loading exceeds the average cross-loading; (2) The names of the categories and items are as follows: Category 1: Leadership; Category 2: Strategy and planning; Category 3: Knowledge and Information; Category 4: People; Category 5: Customer and market focus; Category 6: Innovation, quality and improvement; Category 7: Success and sustainability; Item 1.1: Strategic direction; Item 1.2: Organizational culture; Item 1.3: Leadership throughout the organization; Item1.4: Environmental and community contribution; Item 2.1: Understanding the business environment; Item 2.2: The planning process; Item 2.3:

Development and application of resources; Item 3.1: Collection and interpretation of data and information; Item 3.2: Integration and use of knowledge in decision-making; Item 3.3: Creation and management of knowledge; Item 4.1: Involvement and commitment; Item 4.2: Effectiveness and development; Item 4.3: Health, safety and well-being; Item 5.1: Knowledge of customers and markets; Item 5.2: Customer relationship management; Item 5.3: Customer perception of value; Item 6.1: Innovation process; Item 6.2: Supplier and partner processes; Item 6.3: Management and improvement of processes; Item 6.4: Quality of products and services; Item 7.1: Indicators of success; Item 7.2: Indicators of sustainability

Table 2**:** Loadings and Cross-Loadings for the BCPE Based on Item Scores Secured by the Applicants of the New Zealand Business Excellence Award

Latent variable/ Category Item	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: (1) ΔV is an arbitrary variable, which shows by how much a loading exceeds the average cross-loading; (2) 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 organizational performance; Item 4.2: Information and knowledge management; Item 5.1: Work systems; Item 5.2: Employee learning and motivation; Item 5.3: Employee well-being 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: Organizational effectiveness results; Item 7.6: Leadership and social responsibility results

Table 3**:** Loadings and Cross-Loadings Based on Item Scores Secured by the Applicants Assessed for Prequalification for the Singapore Quality Award

Latent variable/ Category Item	1	2	3	4	5	6	7	Average Cross-loading	**ΔV**
Item 1.1	**0.90**	0.73	0.64	0.65	0.63	0.60	0.61	0.64	0.26
Item 1.2	**0.89**	0.65	0.55	0.70	0.59	0.59	0.59	0.61	0.28
Item 1.3	**0.81**	0.62	0.48	0.58	0.60	0.50	0.64	0.57	0.24
Item 2.1	0.76	**1.00**	0.73	0.69	0.71	0.67	0.64	0.70	0.30
Item 3.1	0.61	0.77	**0.92**	0.68	0.63	0.73	0.60	0.67	0.25
Item 3.2	0.54	0.52	**0.87**	0.51	0.56	0.55	0.57	0.54	0.32
Item 4.1	0.63	0.60	0.57	**0.89**	0.53	0.58	0.65	0.59	0.30
Item 4.2	0.63	0.54	0.55	**0.82**	0.46	0.48	0.60	0.54	0.27
Item 4.3	0.63	0.58	0.62	**0.87**	0.52	0.66	0.62	0.60	0.27
Item 4.4	0.57	0.57	0.53	**0.82**	0.42	0.60	0.57	0.54	0.28
Item 4.5	0.68	0.65	0.61	**0.84**	0.58	0.64	0.68	0.64	0.20
Item 5.1	0.51	0.50	0.38	0.51	**0.61**	0.51	0.32	0.46	0.15
Item 5.2	0.55	0.60	0.59	0.43	**0.80**	0.53	0.65	0.56	0.24

Item 5.3	0.52	0.51	0.52	0.43	**0.84**	0.44	0.51	0.49	0.35
Item 6.1	0.61	0.65	0.68	0.64	0.61	**0.94**	0.56	0.62	0.31
Item 6.2	0.64	0.59	0.64	0.66	0.61	**0.92**	0.59	0.62	0.30
Item 6.3	0.57	0.61	0.71	0.64	0.59	**0.92**	0.52	0.60	0.32
Item 7.1	0.57	0.53	0.57	0.59	0.50	0.55	**0.77**	0.55	0.22
Item 7.2	0.59	0.52	0.53	0.57	0.62	0.46	**0.88**	0.55	0.33
Item 7.3	0.64	0.59	0.60	0.78	0.53	0.57	**0.83**	0.62	0.21
Item 7.4	0.56	0.51	0.51	0.51	0.65	0.44	**0.90**	0.53	0.37
Average ΔV									0.28

Note: (1) ΔV is an arbitrary variable, which shows by how much a loading exceeds the average cross-loading; (2) The names of the categories and items are as follows: Category 1: Leadership; Category 2: Planning; Category 3: Information; Category 4: People; Category 5: Processes; Category 6: Customers; Category 7: Results; Item 1.1: Senior executive leadership; Item 1.2: Organizational culture; Item 1.3: Responsibility to the community and environment; Item 2.1: Strategy Development and deployment; Item 3.1: Management of information; Item 3.2: Comparison and benchmarking; Item 4.1: Human resource planning; Item 4.2: Employee involvement and commitment; Item 4.3: Employee education, training and development; Item 4.4: Employee health and satisfaction; Item 4.5: Employee performance and recognition; Item 5.1: Innovation process; Item 5.2: Process management and improvement; Item 5.3: Supplier and partnering process; Item 6.1: Customer requirements; Item 6.2: Customer relationship 6.3: Customer satisfaction; Item 7.1: Customer results; Item 7.2: Financial and market results; Item 7.3: People results; Item 7.4: Operational results

Ranking of the three BE models on the basis of the level of measurement validity

In order to further analyze how serious the cross-loading issue is, the following were calculated for the three models from the loading and cross-loading data.

(i) Percentage of items that return ΔV values at least as great as 0.20 (ΔV is an arbitrary symbol used by us to resemble 'difference in value'; ΔV for each measurement item was defined as the *loading* minus the *average cross-loading*); the score based on this heuristic was called **Score 1**.

(ii) Percentage of cross-loadings that are equal to, or less than, 0.60; the score based on this heuristic was called **Score 2**.

(iii) Percentage of items that return at least 5 cross-loadings (note that each item has 6 cross-loadings) that are equal or less than 0.60; the score based on this heuristic was called **Score 3.**[3]

The scores reported by the three models based on the above heuristics are shown in Table 4.

Table 4: Performance of the Three Models Based on the Heuristics Used to Test the Relative Level of Validity

Name of the BE Model	**Marks returned** (maximum possible is 100%)		
	Score 1	**Score 2**	**Score 3**
SQAC	95 %	56 %	36%
BCPE	89 %	40 %	21%
ABEF	55 %	13 %	5%

Note that the percentage scores tabulated under 'score 1' in Table 4 refer to the percentage of measurement items that passed a very lenient heuristic on validity. For example, in the case of the SQA criteria, the 'score 1' was 95% because 20 out of the 21 measurement items ((20/21)x100 = 95) met the requirement in the first heuristic. Also note that the second heuristic (Score 2) is more rigorous than the first heuristic (Score 1), while the third heuristic (Score 3) is yet more rigorous than the second. The bottom line is that if stringent validity standards (or rules of thumb) on discriminant validity are applied, all three BE models perform poorly.

It is important to note that the low discriminant validity is a major concern (see Sousa and Voss, 2002 on potential reasons for low discriminant validity of BE models) and it is probably the main reason why strong correlations exist between the seven constructs, to the extent that independent variables in the structural models are nearly multicollinear, in all three BE models. We observe that correlations between the constructs are particularly

[3] Note that the choice of 0.60 as the maximum permissible cross-loading for the heuristics was purely arbitrary. If a value of 0.50 is used (as commonly used in exploratory factor analysis) the marks returned for 'Score 3' for all three models would be 0%, which explains why researchers are lenient with measurement scales developed for new concepts (more about this later).

strong in the case of the ABEF (Table 5); the correlation matrix of the seven constructs of the BCPE and the SQAC are shown in Table 7 and Table 8 in Appendix 2.

Table 5: Correlations of the ABEF Constructs/Latent Variables

Number/Name of the Construct	1	2	3	4	5	6	7
1 (Leadership)							
2 (Strategy and Planning)	0.85						
3 (Knowledge and Information)	0.81	0.81					
4 (People)	0.86	0.78	0.73				
5 (Customer and Market Focus)	0.75	0.82	0.73	0.67			
6 (Innovation, Quality and Improvement)	0.79	0.82	0.85	0.72	0.79		
7 (Success and Sustainability)	0.80	0.76	0.80	0.73	0.65	0.81	
Note: $N = 110$; for all the correlation coefficients $p < 0.001$							

Concerns on near multicollinearity

The near multicollinearity of the constructs causes potentially serious interpretation issues. Foremost, it casts serious doubt as to whether as many as six separate enabler categories (i.e., all seven constructs other than the construct that represents business results/business success) are necessary to conceptualize the leadership and performance management interventions planned and deployed in an organization. The results suggest that it may well be possible to represent all the key management interventions of an organization through a single construct; indeed some researchers have used a single construct previously for this purpose (e.g. Prajogo and Brown, 2004). At a theoretical level though, all the enabler categories are necessary to explain the phenomenon of BE as these constructs are tied to causal propositions on BE. Another issue with multicollinearity is that it undermines the results presented in the structural models (Figures 1 - 3), which will be discussed in the next section (statistical conclusion validity).

Under the above circumstances we believe that it is desirable to refine the measurement items of the three BE models (ABEF in particular). If one were to treat BE constructs as

latent variables possessing strong psychometric properties, then one may be inclined to consider refining any measurement item that returns moderately high cross loadings - say > 0.50 - a rule-of-thumb often cited in exploratory factor analysis (e.g., Hair, Anderson, Tatham, and Black, 1998). If a maximum cross-loading cut-off value of 0.50 were to be used to our set of observations, it appears that all three BE models call for a major overhaul. However, it is cited in literature that such a rule-of-thumb is too rigid for measurement scales (such as BE) that are at an early stage of development (Chin, 1998, pp. 325-326). Thus, considering the loadings and the cross-loading patterns of all three BE models jointly, we recommend that all items that return ΔV values (ΔV is an arbitrary variable, which shows by how much the loading exceeds the average cross-loading for a given measurement item) less than 0.20 and/or a loading less than 0.71 be especially considered for refinement. The ΔV values for each item in each BE model are depicted in Tables 1 - 3. Our recommended criterion calls for refinement of several items in the ABEF. This is obviously reflected in *Score 1* shown in Table 4.

While we believe that our heuristics would be useful in identifying the measurement items that need more attention in future model revisions, we advise the reader to have a flexible approach: our results should always be interpreted in the context of the overall model and its objectives. For example, although $\Delta V < 0.20$ criterion (and/or loading > 0.71) as applied to BCPE (based on New Zealand data) suggest that two measurement items belonging to the Business Results Category in the BCPE need attention, on account of lower ΔV values (Table 2), we can demonstrate through PLS that this discrepancy probably occurred due to attempting to represent all the key stakeholder results through a single construct. We note that in the EFQM Excellence Model (EFQM, 2006), the European counterpart of the BCPE, as many as four constructs are used to represent stakeholder results. The impact of the multidimensionality of the Business Results Category of the BCPE became evident to us when we observed that the ΔV values of all BE items rose noticeably (> 0.20) when a fresh set of PLS output was obtained when item 7.1 was deleted. It is important to note that item 7.1 was deleted merely to observe how the *Business Results* construct behaves when it is modified; however, deleting a measurement item can affect the content validity of a measurement instrument (Nunnally and Bernstein, 1994). Thus, considering the parsimonious nature of the BCPE,

there is probably no need to pay special attention to the aforementioned measurement items in the BCPE by the New Zealand Business Excellence Foundation.

We observed that the PCA analysis yielded loading and cross-loading patterns that were similar (in value) to those presented in Tables 1 - 3. Table 6 (Appendix-1) depicts the PCA results for the SQAC (PCA results for the ABEF and the BCPE are not shown due to space limitations). This suggests that our conclusions on measurement validity on the three BE models are likely to be correct, irrespective of whether or not causal relationships hypothesized exist in the real word. This reinforces our findings on the measurement validity.

Finally, it is important to note that measurement reliability, which is a necessary requirement for validity, was assessed both in terms of coefficient α (Cronbach, 1951) as well as the composite reliability coefficient ρ_c (Werts, Linn, and Jöreskog, 1974) and that all coefficients (i.e., coefficients for each construct in each BE model) easily exceeded the lower bound acceptable value of 0.70 prescribed by Nunnally (1978). These values are not reported due to space limitations.

Statistical conclusion validity

Examination of the structural paths in all three models (Figures 1 - 3) reveals that the majority of these are substantial/significant. These findings compare favorably with prior research (e.g., Flynn and Saladin, 2001; Jayamaha *et al*. 2008; Lee *et al.*, 2003). Moreover, examination of the R^2 values suggest that dependent latent variables (constructs) are well predicted by the independent latent variables, which seems to suggest that all three BE models are useful in a causal-predictive sense.

However, it is important to note that near multicollinearity of the independent variables in each model is cause for concern as it could undermine the results presented in Figures 1 - 3. For example, the statistically nonsignificant path in Figure 1 gives the impression that in the case of the ABEF, *leadership* does not directly relate to *customer focus* (despite the high correlation between the two constructs as shown in Table 5). This is an observation that contradicts a key proposition in the ABEF. It can be shown (using standard equations

used in path analysis) that this is caused by high correlations among three variables, which predict customer focus (i.e., leadership, strategic planning, and information). We believe that with more refined measurement items it is possible to eliminate such confounding situations.

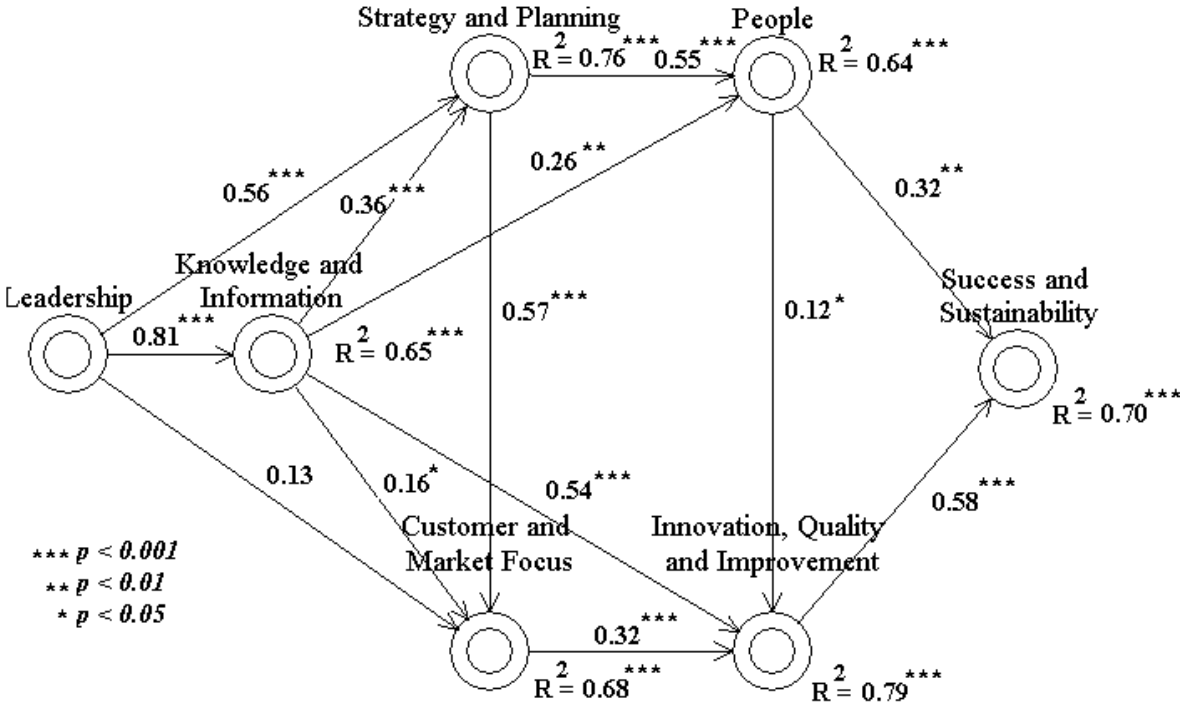

Figure 1: The PLS structural model for the ABEF

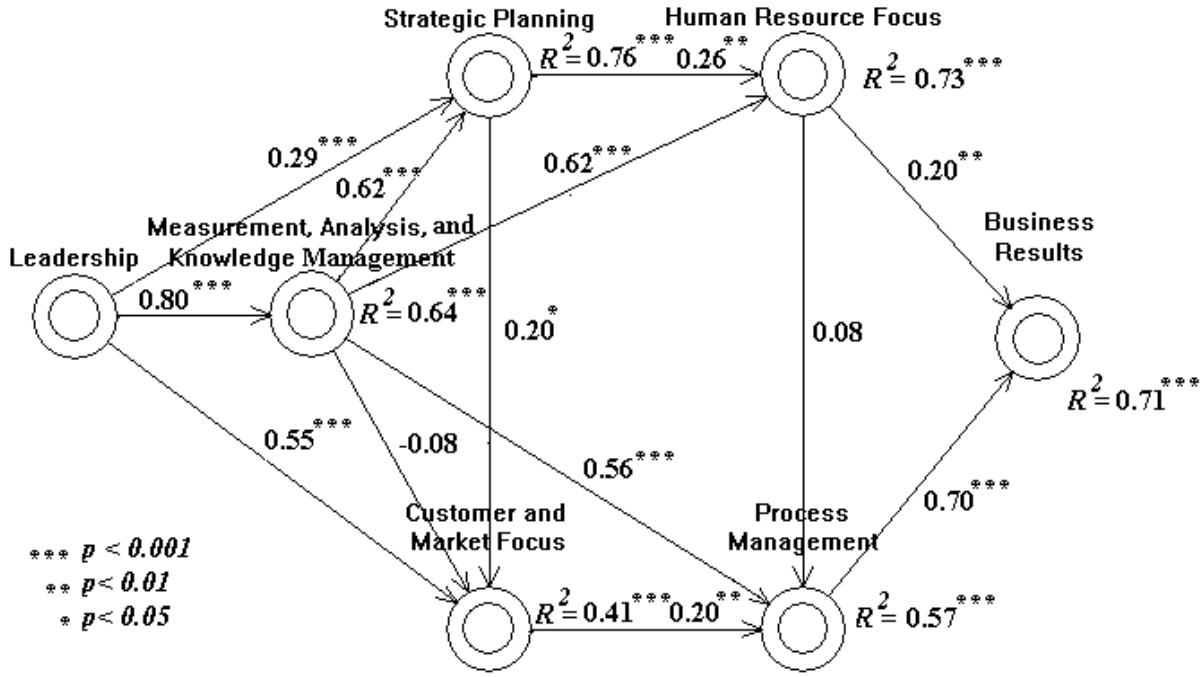

Figure 2: The PLS structural model for the BCPE (New Zealand data)

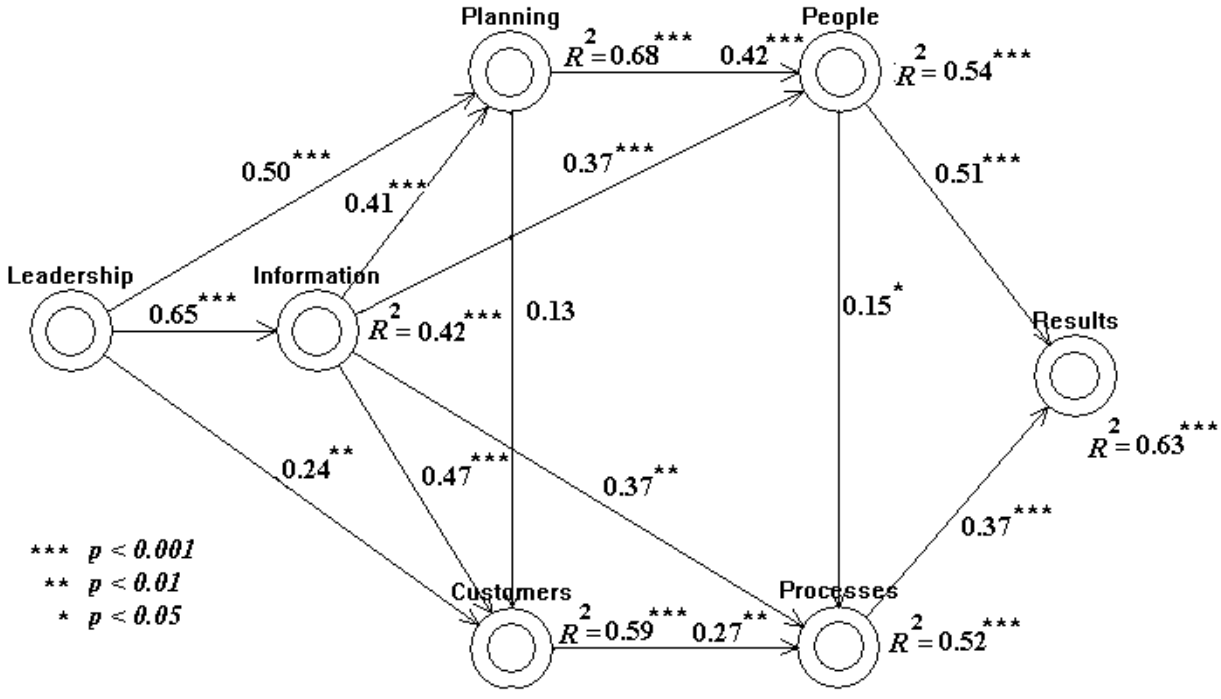

Figure 3: The PLS Structural Model for the SQA Criteria

Part II

In BE, the causal antecedents of business results are *process* and *people* (Figure 3). Thus, in theory, it is possible to predict business results from people and processes (note that the two concepts people and processes can have alternative labels, depending on which BE framework one is referring to, as evidenced in Figure 1 and Figure 2). However, there are three other causal antecedents of processes and people, which in turn are dependent on leadership. Consequently all six enabler categories—as they are justifiably labeled—are determinants of the business results category. The same can be said about the items belonging to the enabler categories (i.e., enabler items). Since we have established the validity of the measurement items[4] used in the three BE models, it makes sense to study the relationships between enablers and results (at item and category levels).

In this section we describe very briefly how we developed our linear regression models to study the relationships between enablers and results and document some of the results we derived and what they imply. Another major objective of this section is to acquaint researchers in quality/BE disciplines about the state-of-the-art multivariate statistical tools that are at their disposal to solve problems involving multicollenearity in regression analysis; we showed earlier that the enabler categories of all three models (ABEF in particular) that we studied, were multicollenear. Statisticians are all too familiar with the problem of multicollenearity of independent variables in traditional (ordinary) least squares regression analysis: unstable regression coefficients (i.e., coefficients that are very much dependent on the dataset they generated them), regression coefficients with unrealistic values (e.g., negative values, when they ought to be positive), inflation of the variances of the regression coefficients (hence often being statistically insignificant, due to small *t*

[4] This is a very important principle in all scientific disciplines. In science any measurement parameter (e.g., time) is based on a theory (e.g., rotation of the earth). Once the scientific community accept the theory to which the parameter is related, they use the parameter routinely to convey meaning. Similarly, once the validity of measures in BE is established, researchers and managers can use these routinely and reliably to convey meaning between the communicators, for performance improvement action.

values),[5] and over fitting—meaning "tailoring the model too much to the current data, to the detriment of future predictions" (Myers, 1990; SAS 2007).

While there are several techniques available to deal with multicollinearity, the partial least squares regression (PLSR) technique is a modern technique that is regarded as the best technique for regression problems involving multicollinearity (Cheng and Wu, 2006; Wold, Ruhe, Wold and Dunn, 1984). Like principal components regression (PCR)—the popular method that was used to handle multicollinearity before PLSR came on to its own—PLSR is a component based procedure, where the objective of the user is to reduce the dimensionality of the variable space prior to regression. This is achieved by selecting only those orthogonal (i.e., uncorrelated) components that are necessary to meet a given criterion (or criteria). This is explained further through an example as follows (for a mathematical treatment and a tutorial on the subject, see Geladi and Kowalski, 1986).

Consider a linear regression problem of finding the regression parameters of the regression model that involves one dependent variable and k independent variables using n observations. If all k orthogonal components (components being derived from the independent variables using PLSR, and each component being a weighted linear sum of the independent variables) are regressed against the dependent variable and the components are subsequently substituted by the independent variables (using the PLSR weights), then one would end up with a set of regression parameters that would be exactly identical to those that could have been derived using ordinary least squares regression (hence such problems as unrealistic parameter values, over-fit or no improvement in fit and so on, if the independent variables are highly correlated). If on the other hand, only one component is regressed against the dependent variable, the results might be unsatisfactory on the grounds that the model is not a good fit to data (i.e. an under-fit). In PLSR, the decision as to how many components ought to be retained for the final analysis is typically based on cross-

[5] The reader can verify some of these peculiarities themselves, using the correlation matrices reported by us (e.g., Table 5).

validation (a more simple method such as observation of incremental R^2 values of the dependent variable against components used is also used as a crude procedure).

STATISTICA 6.0 (the software package used in this study) uses a basic cross-validation procedure known as the *holdout sample method*. In this procedure the user randomly picks part of the cases—say about 80% as we did—for parameter estimation and the balance cases for calculation of the prediction error—more specifically the *Prediction Error Sum of Squares* (PRESS). The optimum model is the model that is derived from that many components that yield a minimum PRESS value. This is accomplished by visual observation of results through a PRESS versus *number of components used plot* (e.g. Figure 4).

Figure 4 depicts the PRESS vs. number of components plot for one of our predictive models. In this model we treated the *Success and Sustainability* category as the dependent variable and the other six categories as the independent variables (obviously the measures relate to the ABEF). In this instance, out of the 110 cases that were available to us, we randomly picked 88 cases (80%)—referred to as the training sample or the analysis sample in statistics—to derive the predictive model; we used the balance 22 cases (20%)—known as the cross-validation sample or the verification sample in statistics—to compute the PRESS. It is clear from Figure 4 that the predictive model formulated from just one component (the first PLSR component) is superior to other five predictive models formulated from multiple components, with addition of each component tending to increase PRESS.[6,7]

---

[6] The PRESS value of 3370 basically means that the sum of squares of the 22 residual values was 3370. This means that the mean residual value of the squares of the residuals is 153.18. The square root of this, which is 12.38, gives the magnitude of the average prediction error. Note that 150 points are allocated to the Success and Sustainability category in ABEF and hence this means that the average perdition error of the model is 12.38 (8%). This suggests that the model is a useful tool in understanding and predicting the results category from the enabler categories.

[7] Note that due to limitations of space, we have not shown PRESS vs. Number of components in respect of all the PLSR models covered in this paper.

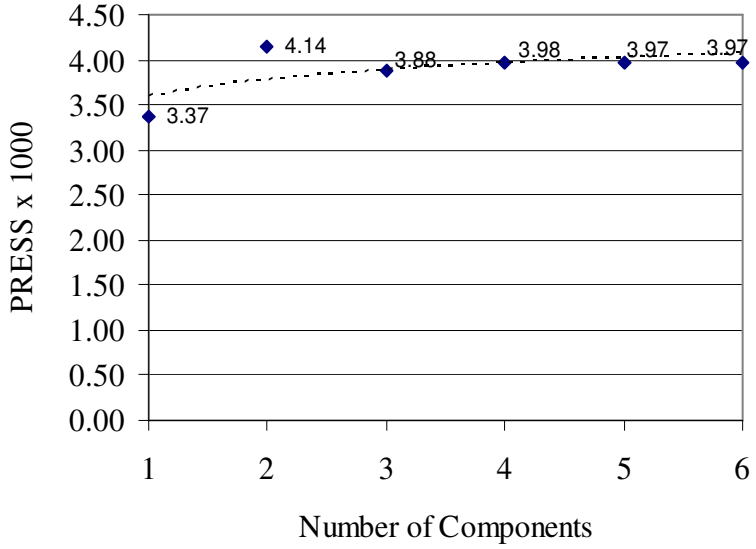

Figure 4: PRESS vs. number of components used (PLSR method) for a predictive model involving the 'Success and Sustainability' category of the ABEF as the dependent variable and the enabler categories as the independent variables

Figure 5 depicts the standardized regression coefficients (β weights) for the PLSR based predictive models involving the results category and the enabler categories for the three BE models. Note that the category names are labeled as per the BCPE and that these labels need to be replaced with the appropriate labels in respect of the other two models (e.g., *Measurement Analysis and Knowledge Management* being replaced by *Knowledge and Information*, in the case of the ABEF). Since we derived the models using the standardized category scores, the regression coefficients shown in Figure 5 are comparable across the three models as well as within the models.

The values reported in Figure 5 imply the following:

- As the range of the beta weights is small (0.1341 to 0.1854), all the enabler categories in all three BE Models are influential in predicting the results categories. It is interesting to note that our study does not suggest that the soft constructs of quality/BE such as leadership and human resource focus are more important than the hard constructs of quality/BE such as process management and information and analysis (in a correlational sense)—as observed by Powell, (1995), Rahman (2001), and Samson and Terziovski (1999).

- The results category of the SQAC is relatively less well predicted by its enabler categories.

- The process management category (and to a lesser extent the strategic planning category) is more influential a predictor of the results category than the other enabler categories. This makes sense, as results are a direct outcome of the processes.

- The effect of the leadership category in predicting business results seem to be approximately the same across all three BE models.

However, we caution that the above findings may not be generalizeable on account of the samples we used (our samples were not random samples).

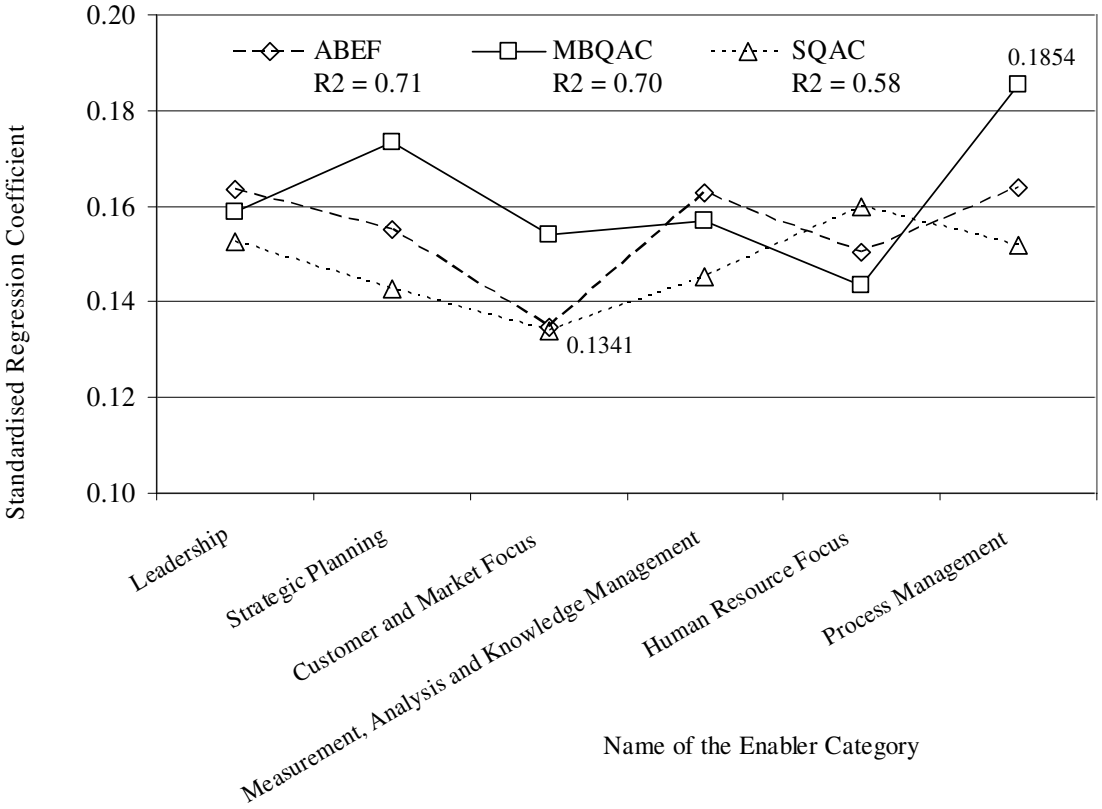

Figure 5: The standardized regression coefficients for the predictive models involving the results category and the enabler categories for the three BE models

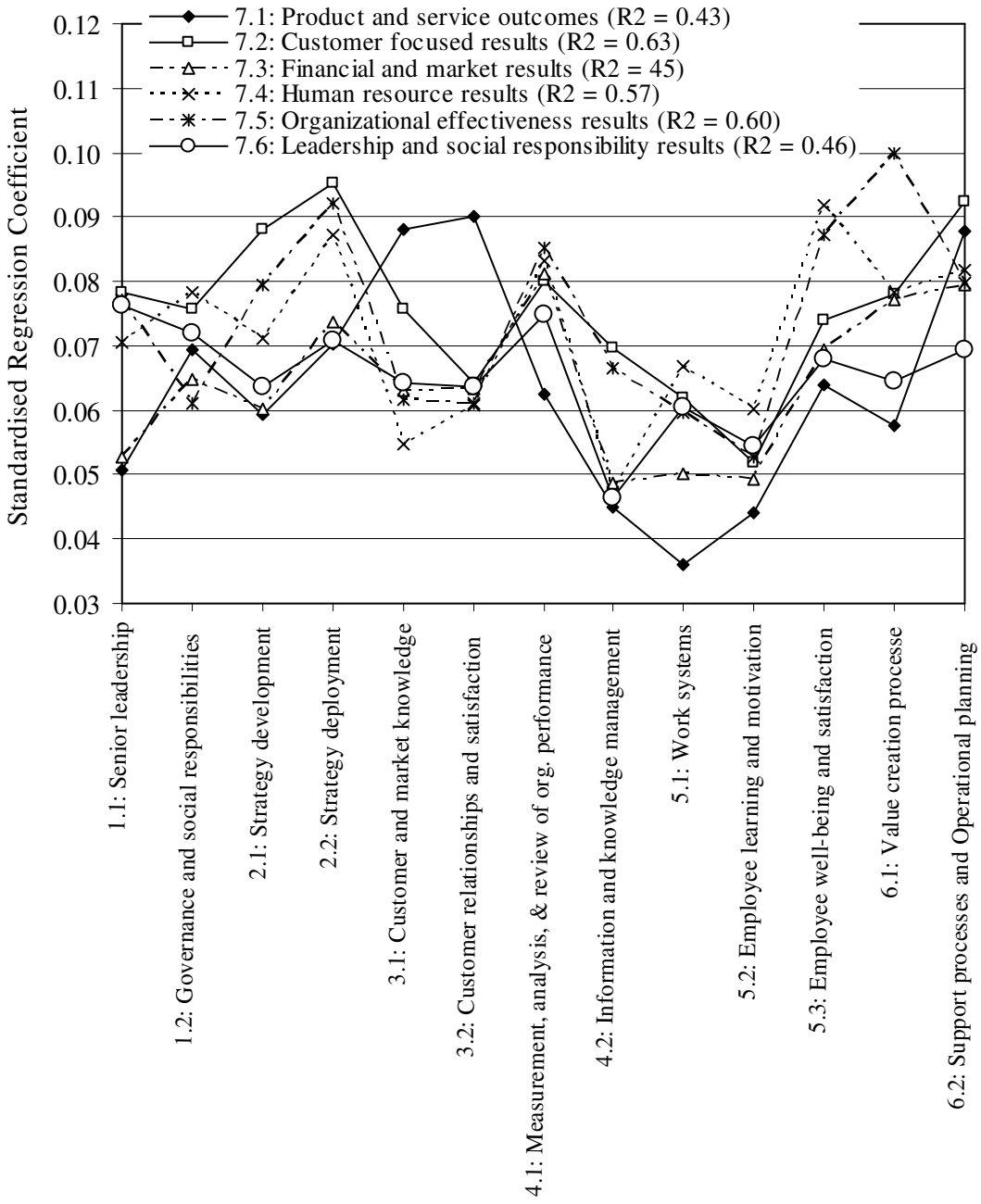

Figure 6: The standardized regression coefficients for the predictive models involving the 'results items' and the 'enabler items' for the BCPE

Figure 6 and Figure 7 depict the standardized regression coefficients for the predictive models involving the enabler items and the results items for the BCPE and the SQAC

respectively. Combining the findings shown in these figures yield the following (again generalization should be made with caution):

1) All enabler items—with perhaps the exception of item 5.1 in SQAC, which was shown to lack validity—are useful predictors of business results.

2) Certain enabler items (as they ought to be) seem to be more influential than others in predicting results for a certain group of stakeholders. For example the enabler items on human resource management (HRM) and human resource development (HRD) are more influential than other enabler items in predicting human resource (or people) results. Similarly, enabler items on process management (ignore item 5.1 of the SQAC) are more influential than other enabler items in predicting operational/organizational effectiveness results.

3) Process management measurement items (items 6.1 and 6.2 in the case of the BCPE and item 5.2 in the case of the SQAC) seem to be more influential than most of the other measurement items in predicting financial and market results (the exceptions being item 4.1 in the case of the BCPE, and item 4.5 in the case of the SQAC). However, as observed by Saunders and Mann (2005), all measurement items appear to exert some influence in predicting financial and market results (as well as results on other stakeholders).

4) Several other enabler items of the BCPE appeared to have influenced more than item 5.1 and item 5.2 in improving the human resources results (item 7.4). This is a bit of a paradox because item 5.1 and item 5.2 are the items that primarily deal with systems and procedures that need to be put in place to enhance intrinsic work motivation of the employees, according to content theories on human motivation (Herzberg, 1987).

The fourth observation mentioned above is worth examining further (ideally using a random sample) to ascertain whether or not this observation is attributable to a sampling error. If not, it may carry connotations peculiar to New Zealand, in which case it is important to investigate what remedial action may have to be put in place to improve the envisaged human resource results.

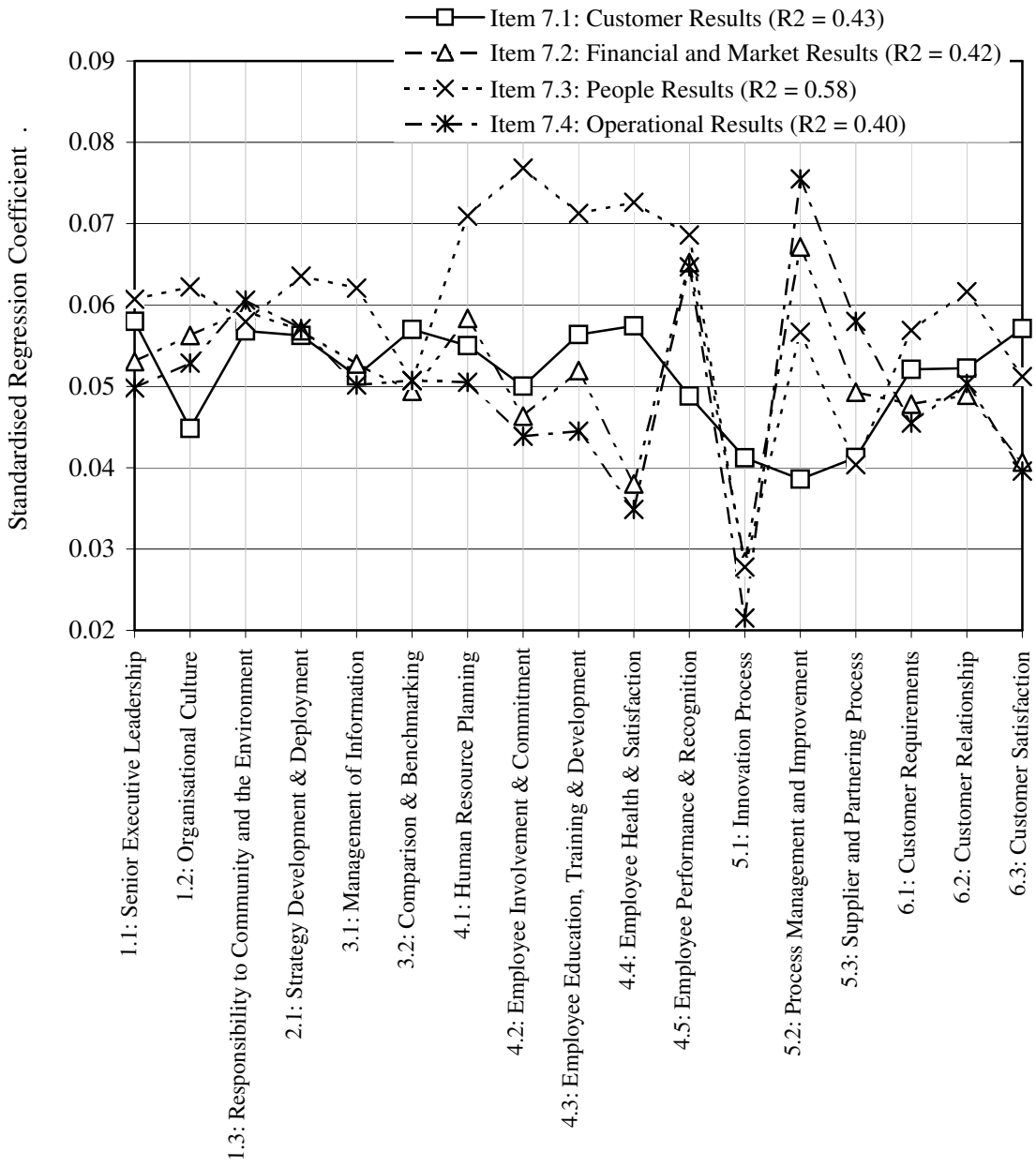

Figure 7: The standardized regression coefficients for the predictive models involving the results items and the enabler items for the SQAC

There are numerous other models that can be developed between the enablers and results; it is not our intention here to depict all of these. As mentioned earlier our intention is to show that notwithstanding the fact that enablers are strongly correlated, with modern statistical

tools, it possible to model the relationships between the enablers and results, if they are reliable and valid. With improved BE measurement scales we believe that PLSR based linear models can be used for predicting and understanding how and to what extent the measures covered by the enabler items impact business results.

Conclusions

Our study is unique in at least three respects: firstly, we used item scores secured by national BE/quality award applicants, to assess the validity of BE models. There are few studies that are based on data secured by BE/quality award applicants. We believe that it is not possible to make an objective assessment of the validity of a national BE model without such data. Proxy methods such as questionnaires/self-reports are more susceptible to various method biases. Secondly, we studied three conceptually analogous models (Australian, New Zealand and Singaporean) within a single theoretical framework using different data sets so that we could enrich our findings by converging evidence. Thirdly, through a state-of-the-art component based regression technique (Partial Least Squares Regression), we developed linear models that have the potential to predict business results and indicate potential performance management problem areas. To our knowledge Partial Least Squares Regression technique (not to be confused with Partial Least Squares path modeling) has not been used before in quality/BE literature.

In Part I of the paper, we demonstrated that all three BE models pass the thresholds for minimum requirements on validity. All three models showed strong evidence of convergent validity, but showed low levels of discriminant validity, which in turn caused a concern about the presence of multicollinearity of the BE categories. This in turn undermined otherwise strong evidence (bearing in mind the fact that BE models are still at the early stage of development) of validity of the three BE models. We devised heuristics to facilitate identifying the measurement items that need more attention in future model revisions. Our heuristics can also be used as a generic method to study measurement validity of instruments that measure concepts that are at early stages of operationalization/scale development. In Part II of the paper, we showed that although certain enabler items seem to

be more influential than others in predicting results (in itself a form of validity[8] known as predictive or pragmatic validity; e.g. Finlay & Wilson, 1997) for a certain group of stakeholders, when results for all the key stakeholders are taken as a whole, all the enabler items and categories become equally influential in predicting the overall results (also see Saunders and Mann, 2005). However, our empirical findings imply that process improvement is relatively more important than most other management interventions, in terms of predicting the financial and market performance of organizations.

Needless to say, generalizations of our findings concerning the validity of the ABEF, BCPE and the SQAC can be made only to the extent that the sample data reflect the population of organizations to which the frameworks apply. As we used non-probability samples—which will always be the case in any empirical study that involves past award applicants—generalization of our findings should be made with caution.

It is desirable that a study similar to ours is carried out with random samples. Clearly, such a study needs the active support of the custodians of BE models, as trained assessors need to be deployed to gather performance measurement scores on various measurement items across a sizable sample in settings outside national quality award administration. In our study, we observed that a more structured measurement approach used in the SQAC did not appear to make those criteria in any way superior to the BCPE (which is relatively less structured) in terms of discriminant validity. Thus, another direction in which future research may head is the conducting of an in-depth study of different major national quality awards as instruments on performance excellence, with a view to proposing how the criteria could be improved to enhance discriminant validity.

[8] See Finlay and Wilson (1997) for a comprehensive list of different types of validities (invariably, all of these validities carry synonyms and quasi-synonyms) that are used in the literature. Note that we have covered all the key types of validities.

Appendix-1

Table 6: Loadings and Cross-Loadings for the Singapore Quality Award (SQA) Criteria based on the PCA method

Latent variable (Category #) Item No.	1	2	3	4	5	6	7	Average Cross-loading	**ΔV**
Item 1.1	**0.90**	0.73	0.64	0.65	0.63	0.60	0.60	0.64	0.26
Item 1.2	**0.89**	0.65	0.55	0.70	0.58	0.59	0.59	0.61	0.28
Item 1.3	**0.82**	0.62	0.48	0.58	0.60	0.50	0.64	0.57	0.25
Item 2.1	0.76	**1.00**	0.72	0.69	0.70	0.67	0.63	0.69	0.31
Item 3.1	0.60	0.77	**0.90**	0.68	0.62	0.73	0.59	0.67	0.23
Item 3.2	0.54	0.52	**0.90**	0.51	0.56	0.55	0.56	0.54	0.35
Item 4.1	0.63	0.60	0.56	**0.89**	0.53	0.58	0.63	0.59	0.31
Item 4.2	0.63	0.54	0.55	**0.82**	0.45	0.48	0.58	0.54	0.28
Item 4.3	0.63	0.58	0.61	**0.88**	0.52	0.66	0.60	0.60	0.28
Item 4.4	0.57	0.57	0.52	**0.83**	0.44	0.60	0.54	0.54	0.29
Item 4.5	0.68	0.65	0.60	**0.83**	0.57	0.64	0.67	0.64	0.19
Item 5.1	0.51	0.50	0.37	0.51	**0.64**	0.51	0.32	0.45	0.19
Item 5.2	0.56	0.60	0.58	0.42	**0.74**	0.53	0.62	0.56	0.20
Item 5.3	0.52	0.51	0.53	0.43	**0.87**	0.44	0.52	0.49	0.38
Item 6.1	0.61	0.65	0.67	0.64	0.60	**0.94**	0.55	0.62	0.32
Item 6.2	0.64	0.59	0.63	0.66	0.59	**0.92**	0.58	0.61	0.31
Item 6.3	0.57	0.61	0.70	0.64	0.59	**0.92**	0.51	0.60	0.32
Item 7.1	0.57	0.53	0.57	0.59	0.50	0.55	**0.77**	0.55	0.22
Item 7.2	0.59	0.52	0.53	0.56	0.60	0.46	**0.89**	0.54	0.35
Item 7.3	0.65	0.59	0.59	0.74	0.52	0.57	**0.81**	0.62	0.20
Item 7.4	0.56	0.51	0.51	0.50	0.63	0.44	**0.91**	0.53	0.39
Average ΔV									0.28

Appendix-2

Table 7: Correlations of the BCPE Constructs/Latent Variables

Number/Name of the Construct	1	2	3	4	5	6	7
1. (Leadership)							
2. (Strategic Planning)	0.79						
3.(Customer and Market Focus)	0.64	0.56					
4. (Meas. Analysis and Knowledge Mgt.)	0.80	0.86	0.52				
5. (Human Resource Focus)	0.81	0.79	0.57	0.84			
6. (Process Management)	0.67	0.73	0.54	0.73	0.67		
7. (Business Results)	0.72	0.79	0.68	0.72	0.67	0.83	
Note: For all the correlation coefficients $p < 0.001$							

Table 8: Correlations of the SQAC Constructs/Latent Variables

Number/Name of the Construct	1	2	3	4	5	6	7
1 (Leadership)							
2 (Planning)	0.76						
3 (Information)	0.65	0.73					
4 (People)	0.74	0.69	0.68				
5 (Processes)	0.70	0.71	0.67	0.59			
6 (Customers)	0.65	0.67	0.73	0.70	0.65		
7 (Results)	0.70	0.64	0.65	0.74	0.68	0.60	
Note: For all the correlation coefficients $p < 0.001$							

Acknowledgements:

The authors are greatly indebted to Darshan Singh and Mak May Yoke of the Standards, Productivity and Innovation Board (SPRING), Singapore, Mary-Anne Bakker and Martin Searle, both formerly of SAI Global, Sydney Australia and Mike Watson and Barbara Nichols of the New Zealand Business Excellence Foundation, Auckland New Zealand, without whose support this project would not have been eventuated.

References

Badri, M. A., Selim, H., Alshare, K., Grandon, E. E., Younis, H., and Abdulla, M. (2006). "The Baldrige Education Criteria for Performance Excellence Framework: Empirical test and validation". *International Journal of Quality and Reliability Management, 23*(9), 1118-1157.

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, 2*(2), 285-309.

Cassel, C. M., Hackl, P., and Westlund, A. H. (2000). "On measurement of intangible assets: a study of robustness of partial least squares". *Total Quality Management, 11*(7), 897-907.

Cheng, B., and Wu, X. (2006). "A modified PLSR method in prediction". *Journal of Data Science, 4*, 257-274.

Chin, W. W. (1998). "The partial least squares approach to structural equation modeling". In G. A. Marcoulides (ed.), *Modern Methods for Business Research* (pp. 295-336): NJ: Lawrence Erlbaum Associates Inc.

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

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

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

Dean, J. W., and Bowen, D. E. (1994). "Management theory and total quality: improving research and practice through theory development". *Academy of Management Review, 19*(3).

EFQM. (2006). *What is Excellence?*. European Foundation for Quality Management, Brussels. http://www.efqm.org/ (accessed March 2007).

Finlay, P. N., & Wilson, J. M. (1997). "Validity of decision support systems: towards a validation methodology. *Systems Research and Behavioral Science, 14*(3), 169-182.

Flynn, B. B., and Saladin, B. (2001). "Further evidence on the validity of the theoretical models underlying the Baldrige criteria". *Journal of Operations Management, 19*(6), 617-652.

Fornell, C., and Cha, J. (1994). Partial Least Squares. In R. P. Bagozzi (Ed.), *Advanced Methods of Marketing Research* (pp. 52-78): Oxford: Blackwell.

Fornell, C., and Larcker, D. F. (1981). "Evaluation Structural Equation Models with Unobservable Variables and Measurement Error". *Journal of Marketing Research, 18*(February), 39-50.

Garvin, D. A. (1991). "How the Baldrige award really works". *Harvard Business Review (Nov.-Dec.), Nov-Dec*, 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, 16*, 91-109.

Geladi, P., and Kowalski, R. B. (1986). "PLS Tutorial". *Analytica Chimica Acta, 185*(1), 1-17.

Grigg, N. P., & Mann, R. S. (2008). "Review of the Australian Business Excellence Framework: A comparison of national strategies for designing, administering and promoting business excellence frameworks". *Total Quality Management and Business Excellence, (forthcoming)*.

Hackman, R. J., and Wageman, R. (1995). "Total Quality Management: Empirical, Conceptual, and Practical Issues". *Administrative Science Quarterly, 40*(2), 309-342.

Hair, J. F., Jr., Anderson, R. E., Tatham, R. L., and Black, W. C. (1998). *Multivariate Data Analysis with Readings* (5 ed.): Englewood Cliffs, NJ: Prentice Hall.

Handfield, R. B., and Ghosh, S. (1995). "An empirical test of linkages between the Baldrige criteria and financial performance". *Proceedings of the Decision Sciences Institute, 3*, 1713-1715.

Hausner, A. (1999). *Business Success and ABEF Evaluation Results: On the nexus between Manufacturing Results and Frameworks for Business Excellence.* Unpublished PhD, University of Wollongong, Wollongong.

Herzberg, F. (1987). "One more time: How do you motivate employees"? *Harvard Business Review, September-October*, 109-120.

Jayamaha, N. P., Grigg, N. P., & Mann, R. S. (2008). "Empirical validity of Baldrige criteria: New Zealand evidence". *International Journal of Quality and Reliability Management, 25*(5), 477-493.

Kanji, G. K. (2002). "Performance measurement system". *Total Quality Management, 13*(5), 715-728.

Kanji, G. K., and Wallace, W. (2000). “Business excellence through customer satisfaction”. *Total Quality Management, 11*(7), 979-998.

Kline, R. B. (1998). *Principles and practice of structural equation modeling*: New York: Guilford Press.

Kristensen, K., and Eskildsen, J. (2006). “Towards a topology on companies striving for organizational excellence”. Paper presented at the 5th Conference of the Multinational Alliance for the Advancement of Organizational Excellence (MAAOE), University of Technology, Sydney, Australia 23 January 2006.

Lee, S. M., Rho, B. H., and Lee, S. G. (2003). “Impact of Malcolm Baldrige National Quality Award Criteria on organizational quality performance”. *International Journal of Production Research, 41*(9), 2003-2020.

MacKenzie, S. B. (2003). “The Dangers of poor construct conceptualization”. *Journal of the Academy of Marketing Science, 31*(3), 323 – 326.

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, 19*(4), 403-426.

Myers, R. H. (1990). *Classical and modern regression with applications* (2 ed.). Boston: PWS-KENT.

NIST. (2006). *Criteria for performance excellence 2006. Baldrige National Quality Program*, from http://baldrige.nist.gov/

Nunnally, J. C. (1978). *Psychometric theory* (2 ed.): New York: McGraw-Hill.

Nunnally, J. C., and Bernstein, I. H. (1994). *Psychometric theory* (3 ed.): New York: McGraw-Hill.

Pannirselvam, G. P., and Ferguson, L. A. (2001). “A study of the relationships between the Baldrige categories”. *The International Journal of Quality and Reliability Management, 18*(1), 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, 16*, 529-550.

Podsakoff, P. M., MacKenzie, S. B., Lee, J.Y., and Podsakoff, N. P. (2003). “Common method biases in behavioral research: A critical review of the literature and recommended remedies”. *Journal of Applied Psychology, 88*(5), 879-903.

Powell, T. C. (1995). “Total quality management as competitive advantage: A review and empirical study”. *Strategic Management Journal, 16*(1), 15-37.

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". *Quality Management Journal*, 11(4), 31-42.

Rahman, S. (2001). "Total quality management practices and business outcome: Evidence from small and medium enterprises in Western Australia". *Total Quality Management, 12*(2), 201-210.

Rosa, M. J. P. D., Saraiva, P. M., and Diz, H. (2003). "Excellence in Portuguese higher education institutions". *Total Quality Management, 14*(2), 189-197.

SAI Global. (2004). *Case Studies in Organizational Excellence 2001-2003: Corrections Victoria* (Case Studies). Sydney: SAI Global Limited.

Samson, D., and Terziovski, M. (1999). "The relationship between total quality management practices and operational performance". *Journal of Operations Management, 17*(4), 393-409.

SAS. (2007). *Partial Least Squares*. from http://support.sas.com/rnd/app/da/new/dapls.html (accessed 28 May 2007),

Saunders, M., & Mann, R. (2005). "Self-assessment in a multi-organizational network". *International Journal of Quality & Reliability Management, 22*(6), 554-571.

Sitkin, S. B., Sutcliffe, K. M., and Schroeder, R. G. (1994). "Distinguishing Control From Learning in Total Quality Management: A Contingency Perspective". *Academy of Management Review, 19*(3), 537-564.

Sousa, R., & Voss, C. A. (2002). "Quality management re-visited: A reflective review and agenda for future research". *Journal of Operations Management, 20*(1), 91-109.

SPRING. (2007). *Singapore Quality Award*, Standards, Productivity and Innovation Board of Singapore. www.spring.gov.sg (accessed 03 March 2007)

Straub, D., Boudreau, M., and Gefen, D. (2004). "Validation guidelines for IS positivist research". *Communications of the Association for Information Systems, 13*, 380-427.

Werts, C. E., Linn, R. L., and Jöreskog, K. G. (1974). "Interclass reliability estimates: Testing structural assumptions". *Educational and Psychological Measurement, 34*, 23-33.

Wilson, D. D., and Collier, D. A. (2000). "An empirical investigation of the Malcolm Baldrige National Quality Award causal model". *Decision Sciences, 31*(2), pp 361-390.

Winn, B. A., and Cameron, K. S. (1998). "Organizational quality: An examination of the Malcolm Baldrige national quality framework". *Research in Higher Education, 39*(5), 491-512.

Wold, H. (1980). “Model construction and evaluation when theoretical knowledge is scarce: Theory and application of partial least squares”, in J. Kmenta and J. D. Ramsey (eds.), *Evaluation of econometric models* (pp. (47-74): New York: Academic Press.

Wold, S., Ruhe, A., Wold, H., and Dunn, W. J. (1984). “The collinearity problem in linear regression: The partial least squares (PLS) approach to generalized inverses”. *SIAM Journal of Scientific and Statistical Computing, 5*(3), 735-742.

Author biographies

Nihal Jayamaha is a final year PhD candidate at the Centre for Organisational Excellence Research (COER), Massey University, New Zealand. His research interests include performance measurement, statistical modelling, and operations research. Nihal has worked previously (over 20 years) in large electrical utilities in South Asia and the Middle East in the capacity of an operations engineer, a technical auditor and a project manager. He holds Bachelors and Masters degrees in Engineering and as well as a Masters in Business Administration.

Dr Robin Mann is the Head of the Centre for Organisational Excellence Research, New Zealand, www.coer.org.nz, Chairman of the Global Benchmarking Network, www.globalbenchmarking.org, Advisory Board member of the e-TQM College, Dubai, www.etqm.ae, and Co-Founder of BPIR.com Ltd, www.bpir.com – a leading benchmarking website resource with over 6,000 members worldwide. Robin’s 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) and leading TRADE benchmarking projects in Singapore (2007 onwards). Robin worked in Edinburgh (1992-1995) for Burton’s Biscuits as a process improvement manager and obtained his PhD in TQM at Liverpool University in 1992.

Dr Nigel Grigg is Senior Lecturer in Quality and Statistics at Massey University in New Zealand, and the coordinator of Massey's suite of postgraduate programmes in Quality Systems. He is a Chartered Quality Professional, member of the Chartered Quality Institute, the American Society for Quality and the NZ Institute of Directors, and a Director of the New Zealand Organisation for Quality. His research interests include process improvement and the use of statistical thinking and methods within the knowledge development cycle. He has authored and co-authored over 80 journal and conference papers on aspects of quality management, and has received three national best journal paper awards, in addition to the 2004 annual National Award from the Institute of Quality Assurance.