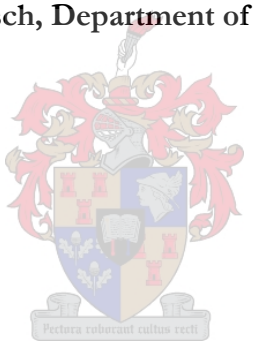
AN INVESTIGATION INTO THE MEASUREMENT INVARIANCE OF THE PERFORMANCE INDEX.

Hazel Dunbar-Isaacson

University of Stellenbosch, Department of Industrial Psychology



Thesis presented in partial fulfilment of the requirements for the degree of Master of Commerce at the University of Stellenbosch

Supervisor

Prof CC Theron

December 2006

ABSTRACT

The leadership-for-performance framework designed by Spangenberg and Theron (2004) aspires to explicate the structural relationships existing between leader competency potential, leadership competencies, leadership outcomes and the dimensions of organizational unit performance. The Performance Index (PI) and Leadership Behaviour Inventory (LBI) comprise the leadership-for-performance range of measures. The PI was developed as a comprehensive criterion measure of unit performance for which the unit leader could be held responsible. The basic PI structural model has been developed to explain the manner in which the various latent leadership dimensions measured by the LBI affect the eight unit performance latent variables that are assessed by the PI. Although preliminary research suggests the basic PI structural model could be refined, continued research in this regard can only be justified if the basic PI measurement model is shown to be measurement invariant across independent samples from the target population. As part of ongoing research of the leadership-for-performance range of measures, this cross-validation study investigated the extent to which the PI measurement model may be considered measurement invariant across two independent samples from the same population. Two samples were collected through non-probability sampling procedures and included 277 and 375 complete cases after imputation by matching. Item analysis and dimensionality analysis were performed on each of the PI sub-scales prior to the formation of item parcels. No items were excluded based on item- and dimensionality analysis results. Two composite indicator variables (item parcels) were created from the items of each sub-scale and were treated as continuous variables in the subsequent statistical analyses. Structural equation modelling, using robust maximum likelihood estimation, was used to perform a confirmatory first-order factor analysis on the item parcels for each sample. The measurement model was fitted to both samples independently and close fit for each sample was established. The measurement model was cross-validated using a progressive series of measurement invariance tests. Results indicated the PI measurement model did not display full measurement invariance across the two samples although it did cross-validate successfully under the configural invariance condition. Statistically significant non-equivalence was found to exist in both the measurement error variances and the factor covariances ($p<0,05$), although the $p<0,05$ critical value was only narrowly surpassed in both cases. The measurement model did, however, display metric invariance across the samples as no significant differences were found between the factor loadings, suggesting the content of each item is perceived and interpreted in a similar manner across samples from the target population. When considered in combination, these results may be viewed as quite satisfactory as they indicate that the

measurement model does not appear to vary greatly when fitted to data from the two samples. As this study has established at least metric invariance of the PI, it therefore provides some basis of confidence for proceeding with subsequent research aimed at establishing the structural invariance of the basic PI structural model and eventually research that links the leadership behaviour to work unit performance as measured by the PI. Limitations of this study are discussed.

OPSOMMING

Die leierskap-prestasieraamwerk daargestel deur Spangenberg en Theron (2004) het as doel om die strukturele verwantskappe wat tussen leierskapbevoegdheidspotensiaal, leierskapbevoegdhede, leierskapuitkomste en die dimensies van organisatoriese eenheidprestasie bestaan eksplisiet te maak. Die Performance Index (PI) en die Leadership Behaviour Inventory (LBI) verteenwoordig die huidige leierskap-gerig-op-prestasiemeetinstrumentreeks Die PI is ontwikkel as 'n omvattende kriteriummeting van organisatoriese prestasie waarvoor die leier van die eenheid aanspreeklik gehou sou kon word. Die oogmerk met die ontwikkeling van die basiese PI strukturele model is om die wyse waarop die onderskeie latente leierskapdimensies wat deur die LBI gemeet word die agt organisatoriese eenheidsprestasiedimensies wat deur die PI gemeet word, affekteer. Ofskoon voorlopige navorsing daarop dui dat die basiese PI strukturele model verfyn sou kon word, sou voortgesette navorsing in hierdie verband slegs geregverdig kon word indien die metingsinvariansie van die basiese PI metingsmodel oor onafhanklike steekproewe uit die teikenpopulasie aangetoon sou kon word. As deel van die voorgesette navorsing op die leierskap-vir-prestasieprodukreeks ondersoek hierdie kruisvalidasiestudie die mate waartoe die PI metingsmodel as metingsinvariant beskou kan word oor onafhanklike steekproewe uit dieselfde populasie. Twee steekproewe is versamel deur middel van nie-waarskynlikheidsteekproefnemingsprosedures en het 277 en 375 waarnemings ingesluit na imputasie deur middel van afparing. Itemontleding en dimensionaliteitontleding is op elk van die PI subskale uitgevoer voor die vorming van itempakkies. Geen items is op grond van die item-en dimensionaliteitontledingsresultate geëlimineer nie. Twee saamgestelde waargenome veranderlikes (itempakkies) is uit die items van elke subskaal bereken en is as deurlopende veranderlikes in die daaropvolgende statistiese ontledings hanteer. Strukturele modellering is met behulp van maksimumaanneemlikheidskattingstegnieke gebruik om 'n bevestigende faktorontleding op die itempakkies op elk van die steekproewe uit te voer. Die metingsmodel is onafhanklik op die twee steekproewe gepas en nou passing is vir elk van die steekproewe gevind. Die metingsmodel is vervolgens gekruisvalideer deur 'n reeks opeenvolgende metingsinvariansietoetse. Die resultate het aangetoon dat die PI metingsmodel nie volle metingsinvariansie oor die twee steekproewe toon nie ofskoon dit wel suksesvol onder die konfigurale-invariansietoestand suksesvol gekruisvalideer het. Statisties beduidende gebrek aan ekwivalensie ($p<0,05$) is gevind in beide die metingsfoutvariansies en die faktorkovariansies, ofskoon die $p<0,05$ kritieke waarde in beide gevalle slegs nouliks oorskrei is. Die metingsmodel het egter metriese-ekwivalensie oor die twee steekproewe getoon insoverre geen beduidende versille in faktorladings oor steekproewe gevind

is nie. Dit impliseer dat die inhoud van die items eenders waargeneem en geïnterpreteer is oor die twee steekproewe uit die tekenpopulasie. Wanneer die resultate in in kombinasie beoordeel word is die gevolgtrekking heel bevredigend insoverre dit daarop dui dat daar nie groot versille bestaan wanneer die metingsmodel op die data van die twee steekproewe gepas word nie. Insoverre hierdie studie ten minste die metriese ekwivalensie van die PI aangetoon het baan hierdie studie die weg om voort te gaan met navorsing gerig op die strukturele ekwivalensie van die basiese PI strukturele model en uiteindelik dan ook navorsing gerig op die koppeling tussen leierskapgedrag en organisatoriese eenheidprestasie soos gemeet deur die PI Beperkinge waaraan die studie onderworpe is word bespreek.

TABLE OF CONTENTS

CHAPTER 3: RESEARCH METHODOLOGY AND PREPARATORY DATA ANALYSES **21**

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

ACKNOWLEDGEMENTS

I am most grateful to Professor Callie Theron for providing me with the opportunity to conduct research into the Leadership-for-performance instruments, and for the patient, diligent and empowering approach to supervising this research. I am also indebted to Professor Herman Spangenberg and Frik Landman, the CEO of USB-ED, for providing assistance with regards to securing participants for this research. Professor Spangenberg, your encouragement during this process made it a more positive experience. I would like to thank my editor and sister, Dale, for kindly assisting in making this paper grammatically sound. Lastly, I would like to thank my husband, Steve Isaacson, and my parents, Stewart and Yvonne Dunbar, for their endless love and support as I have worked towards getting my professional qualification.

CHAPTER 1
ON THE NEED FOR A CROSS-VALIDATED COMPREHENSIVE UNIT PERFORMANCE MEASURE

1.1 INTRODUCTION

To meet the challenge of sustained competitiveness and profitability in the context of immense international and domestic competition and change, organisations are increasingly focusing on the extent to which leaders are able to positively influence the performance of their individual followers and work units (Bass, Avolio, Jung & Berson, 2003; Bunderson & Sutcliffe, 2003; House, 1998; Kolb, 1996; Yukl, 2002). This realisation has led researchers and practitioners to pay more attention to the relationship between leaders' behavioural competencies and individual or work unit performance. As the importance of contributions of work unit performance towards organisational performance has become increasingly acknowledged, so has the need to effectively measure work unit performance become a reality.

This chapter discusses the need for a comprehensive work unit performance measure and provides examples of measures of work unit performance used in recent research. The development and underlying structure of the Performance Index (PI) is discussed in relation to this need.

1.2 UNIT PERFORMANCE MEASURES: THE NEED FOR A COMPREHENSIVE MEASURE

Most research on workplace effectiveness has historically focused on performance outcomes at the individual employee level and comparatively less is known about work unit performance and its antecedents (Gelade & Ivery, 2003). Although individual effectiveness is undoubtedly an essential component of superior work unit performance, many types of organisational behaviour (e.g. climate) and many indicators of organisational performance are more relevant to the work unit. In addition, typical traditional measures of work unit performance characteristically fall short of what is required of today's measures as they do not encompass all the performance dimensions for which the unit leader should be held accountable (Green, Madjidi, Dudley & Gehlen, 2001; Sale & Inman, 2003; Spangenberg & Theron, 2004).

In contrast, effective measures of performance should enable an organisation to identify what to improve on, and how best to use its limited resources more effectively in order to facilitate this improvement (Kanji, 2002). Traditional measures fall short of this criteria in three ways. Firstly, traditional measures almost exclusively focus on financial measures which tend to reflect the consequences of decisions, sometimes well after decisions have been made. Thus, they are generally considered to be lagging indicators that have little predictive power. Secondly, traditional measures tend to focus on outcomes, rather than processes that are at the core of management. Process management requires more transversal measures, which traditional systems do not provide. Thirdly, traditional measures are seen to promote a short-term vision and the overemphasis of conforming to conventional standards rather than seeking innovative solutions (Kanji, 2002).

1.3 EXAMPLES OF MEASURES OF UNIT PERFORMANCE

In a review of research in which measures of unit performance were employed, substantially fewer researchers used a more balanced approach to measuring unit performance that included both financial and non-financial measures, rather than only traditional financial measures (Sale & Inman, 2003). Examples of traditional measures of work unit performance identified in recent research include: "the bosses' perception of whether the unit was performing above average compared to other units reporting to the boss" (Javidan & Waldman, 2003, p. 231); profitability relative to targets or units sold (Avolio, Howell & Sosik, 1999; cited in Safferstone 1999, p. 103; Bunderson & Sutcliffe, 2002) or simply net operating profits before tax (Bunderson & Sutcliffe, 2003).

Similarly, in the non-profit context where the need for non-financial measures of performance is apparent, the effectiveness of non-profit organisations has proved a difficult concept to define and operationalise, although such organisations exist to render a public service and logically their effectiveness should be measured by how well they perform this service, and not only by financial performance. Nonetheless, not-for-profit performance measures that are traditionally used tend to mirror that of for-profit organisations as they also focus on medium to short-term goal achievement, and have been criticised for the lack of emphasis placed on evaluating processes used to attain these goals that would sustain performance (Green *et al.,* 2001).

In comparison, good performance measures should cover a broad spectrum of measures and provide data not only on financial success, but also an organisation's strategic issues, such as quality, responsiveness and flexibility. They should include multiple measures in order to avoid misleading interpretations resulting from the use of a single dimension or a narrow definition of performance (Sale & Inman, 2003). Such measures should also achieve compatibility and integration and align core business processes, and be valid, reliable and easy to use (Kanji, 2002). A reason why traditional measures continue to be used so widely was proposed by Sale and Inman (2003) who recognised that although research indicates that pay is increasingly linked to a business unit or organisation's ability to respond to its competitive realities, performance incentives (for example, gain sharing, profit sharing, and productivity dividends) are typically based on traditional financial performance measures.

Although measures of work unit performance are far from perfect, the situation is not entirely one-sided. In a review of performance measurement research, Forbes (1998) recognized that non-conventional, 'emergent' approaches to measuring effectiveness were increasingly being used. An example that supports Forbes' analysis includes Globerson and Riggs' (1989) paper which called for measuring unit performance along several dimensions. Globerson and Riggs (1989) promoted the use of operational performance criteria in addition to traditional financial measures that would allow managers to make better short-term operating decisions that promote long-term organisational productivity. They included five types of operational measures in a matrix proposed for developing performance objectives and indicators, namely: (a) output quantity, (b) resource utilization, (c) operating efficiency, (d) quality and timeliness, (e) employee behaviour.

A further example is the introduction of The Balanced Scorecard (BSC) by Kaplan and Norton (1992) that is considered by some as a great step forward towards overcoming some of the limitations of the traditional financial measures, and is widely used by businesses and therefore deserves a mention in this discussion (Lipe & Salterio, 2000). The BSC includes financial measures, customer relations, internal business processes and organisational learning and growth activities (Kaplan & Norton, 1996a). The BSC is fairly complex and relatively costly to develop and implement as it needs to be tailored to each unit's goals and strategies, and allows for specific indicators to be chosen for each individual business unit. However, the BSC is at a disadvantage in circumstances in which a generic, standardized work unit performance measure is required to

compare many leaders' behaviours to their work unit's performance for the purpose of improving leader effectiveness and ultimately unit performance.

Recent research by Loughry (2002), Fay, Luhrmann and Kohl (2004), Gibson and Birkinshaw (2004), Gelade and Ivery (2003), and Watson and Wooldridge (2005) also support Forbes' (1998) claim that more balanced approaches to measuring unit effectiveness are being employed. Loughry (2002) used two measures of performance to examine the relationship between peer monitoring levels in work units and the work units' performance. Overall unit performance was measured through the manager's evaluation of the units' overall performance including speed of service, guest courtesy, quantity of work, quality of work, cleanliness of area, teamwork and value of services provided by the unit. Problem-free performance included managers' evaluation of the degree to which the work units were free of employee behaviour problems such as absenteeism, tardiness, disciplinary problems, mistakes and accidents, and employee bickering.

Similarly, Fay *et al.* (2004) asked managers to rate their units' performance on six specific performance aspects, using 5-point Likert type scales with 5 referring to what the managers perceived to be a "very good result" and 1 referring to a "very poor result". Three aspects referred to performance regarding the effectiveness of business processes: time wasted on process barriers; speed of core business processes; productivity; and three aspects assessed the financial side of centre performance: profits, business volume, and deviations from planned budgets. A study by Gelade and Ivery (2003) evaluated the effectiveness of human resource management on performance. For this research a composite measure of overall unit performance was computed by averaging the standardized scores for sales against target, customer satisfaction, staff retention, and clerical accuracy.

Gibson and Birkinshaw (2004) measured unit performance using four items that required senior and middle management respondents to reflect on work unit performance over the last five years and indicate the degree to which they agreed with the following: (1) "This business unit is achieving its full potential", (2) "People at my level are satisfied with the level of business unit performance", (3) "This business unit does a good job of satisfying our customers", and (4) "This business unit gives me the opportunity and encouragement to do the best work I am capable of". An external validity check was conducted on this performance measure by comparing it to financial performance indicators, including return on assets (ROA), return on equity (ROE), and shareholder return over a five-year period for each company. The measures of financial performance were found to be highly correlated with aggregated measures of subjective

performance as rated by senior managers, lending strong external validity to the subjective performance measure.

In order to examine the influence of business unit managers on corporate-level strategy formulation, Watson and Wooldridge (2005) used a questionnaire formulated by Gupta and Govindarajan (1986, cited in Watson and Wooldridge, 2005, p. 148) to measure work unit performance. This measure provides a weighted average of business unit performance from the following two questions: (1) How *important* is each of the following dimensions of the performance to your organization: (a) return on investment, (b) profit, (c) cash flow from operations, (d) cost control, (e) development of new products, (f) sales volume, (g) market share, (h) market development, (i) personnel development, and (j) political-public affairs? (2) How *effective* is your organization on each of the following dimensions of performance: (a) return on investment, (b) profit, (c) cash flow from operations, (d) cost control, (e) development of new products, (f) sales volume, (g) market share, (h) market development, (i) personnel development, and (j) political-public affairs? For each dimension, respondents were asked to indicate the performance of the business unit relative to its industry competitors on a seven-point scale, and the importance of the dimension on a five-point scale.

The above examples highlight similarities and shortcomings relating to how work unit performance is currently measured. Firstly, the lack of consensus as to what a measure of work unit performance should include is quite apparent as almost all measures differed substantively. Secondly, traditional measures of unit performance which are typically lagging measures continue to be used in isolation, whereas a more balanced approach that includes both financial and non-financial measures would allow researchers increased confidence in their research findings. By far the most comprehensive measure used in recent research appears to be Gupta and Govindarajans' (Watson and Wooldridge, 2005) measure which was originally been designed for a study that researched resource sharing among business units. However, this measure does not appear to have been used extensively as no other reference to it could be found in the literature survey, and there is no information on the theoretical model or validity and reliability of the measure. Lastly, most of the measures are highly subjective. Only the study of Gibson and Birkinshaw (2004) established the external validity of their performance measure. The above examples of recent research that included measures of work unit performance support the conclusion by Spangenberg and Theron (2002) that there is no generic standardized measure of work unit performance that can serve as a criterion measure of work unit performance.

1.4 THE DEVELOPMENT AND UNDERLYING STRUCTURE OF THE PERFORMANCE INDEX

This above mentioned shortage of generic and standardized measures of work unit performance that could serve as a criterion variable became apparent to Spangenberg and Theron (2002; 2004) when they needed to validate the Performance Management Audit Questionnaire, (Spangenberg & Theron, 1997) and more recently in the design of the Leadership Behaviour Inventory (LBI). The LBI is a comprehensive leadership questionnaire that serves to identify latent leadership dimensions on which a leader under-performs. The LBI forms one component of Spangenberg and Theron's (2004) envisaged leadership-for-performance framework. The leadership-for-performance framework aspires to explicate the structural relationships existing between leader competency potential, leadership competencies, leadership outcomes and the dimensions of unit performance (Theron, Spangenberg & Henning, 2004). To develop and evaluate this framework, a comprehensive conceptualization of organizational work unit performance and a corresponding performance measure that could be used in conjunction with the LBI were required.

In their review of available measures, Spangenberg and Theron (2004) identified two psychometric measures of organisational performance that were applicable to their needs, namely Nicholson and Brenner's (1994) dimensions of perceived organisational performance, and the Unit Performance Questionnaire (UPQ) (Cockerill, Shroder & Hunt, 1993, cited in Spangenberg & Theron, 2004, p. 19). As with the examples referred to in the above chapter though, neither of these performance measures covered the unit performance domain comprehensively enough to successfully serve the purpose of a work unit criterion measure (Spangenberg & Theron, 2004). In response to this need Spangenberg and Theron (2004) developed a generic, standardized unit performance measure, the Performance Index (PI), which encompasses the unit performance dimensions for which the unit leader could be held responsible.

The PI was built on a comprehensive structural model of work unit performance effectiveness that was based on literature targeting financial and non-financial performance measures of organisational effectiveness (Spangenberg and Theron, 2004). The resulting PI model is a synthesis of Nicholson and Brenner's (1994) systems approach, Conger and Kanungo's leadership outcomes (Conger and Kanungo, 1998), and Gibson, Ivancevich and Donelly's (1991) time-dimension model of organisational performance. The final version of the Performance Index questionnaire includes 56 questions which cover eight latent dimensions of unit

performance. The dimensions, with a brief description of each dimension, are presented in Table 1.

TABLE 1

BRIEF SUMMARIES OF THE PI UNIT PERFORMANCE DIMENSIONS

1. Production and efficiency	Refers to quantitative outputs such as meeting goals, quantity, quality and cost-effectiveness, and task performance.
2. Core people processes	Reflect organisational effectiveness criteria such as goals and work plans, communication, organisational interaction, conflict management, productive clashing of ideas, integrity and uniqueness of the individual or group, learning through feedback and rewarding performance.
3. Work unit climate	Refers to the psychological environment of the unit, and gives an overall assessment of the integration, commitment and cohesion of the unit. It includes working atmosphere, teamwork, work group cohesion, agreement on core values and consensus regarding the vision, achievement-related attitudes and behaviours and commitment to the unit.
4. Employee satisfaction	Considers individual's satisfaction with the task and work context, empowerment, and career progress, as well as with outcomes of leadership, e.g. trust in and respect for the leader and acceptance of the leader's influence.
5. Adaptability	Reflects the flexibility of the unit's management and administrative systems, core processes and structures, capability to develop new products/services and versatility of staff and technology. It reflects the capacity of the unit to respond appropriately and expeditiously to change.
6. Capacity (wealth of resources)	Reflects the internal strength of the unit, including financial resources, profits and investment, physical assets and materials supply and quality and diversity of staff.
7. Market share/scope/ standing	Includes market share (if applicable), competitiveness and market-directed diversity of products or services, customer satisfaction and reputation for adding value to the organisation.
8. Future growth	Serves as an overall index of projected future performance and includes profits and market share (if applicable), capital investment, staff levels and expansion of the unit.

(Theron, Spangenberg & Henning, 2004, p. 36)

The PI uses a Likert-type scale with descriptive responses ranging from 1 (describes poor performance for the item) to 5 (describes exceptional performance for the item). Respondents may select a non-observable rating as a last resort if they believe that they are not in a position to accurately evaluate the work unit on the particular dimension.

1.5 PREVIOUS RESEARCH ON THE PERFORMANCE INDEX MODEL FIT

As a comprehensive criterion measure of unit performance, the PI model is intended to explain the manner in which the various latent leadership dimensions measured by the LBI affect the eight unit performance latent variables that are assessed by the PI. Before such research may be conducted the PI should be cross-validated across samples of the target population. Conducting cross-validation research would, however, not be appropriate without the foundation of prior research by Henning, Theron and Spangenberg (2003) and Theron, Spangenberg & Henning (2004) which investigated the internal structure of the PI.

In their initial study, Henning *et al. (*2003) suggested hypotheses on the inter-relationships between the eight unit performance latent variables described above. Confirmatory factor analysis was performed and the results indicated satisfactory factor loadings on the measurement model which supported acceptable measurement model fit. The proposed structural model of the PI was also found to have good fit and these initial findings suggested that the eight dimensions of the PI model should be seen to influencing each other as illustrated in the structural model in Figure 1.

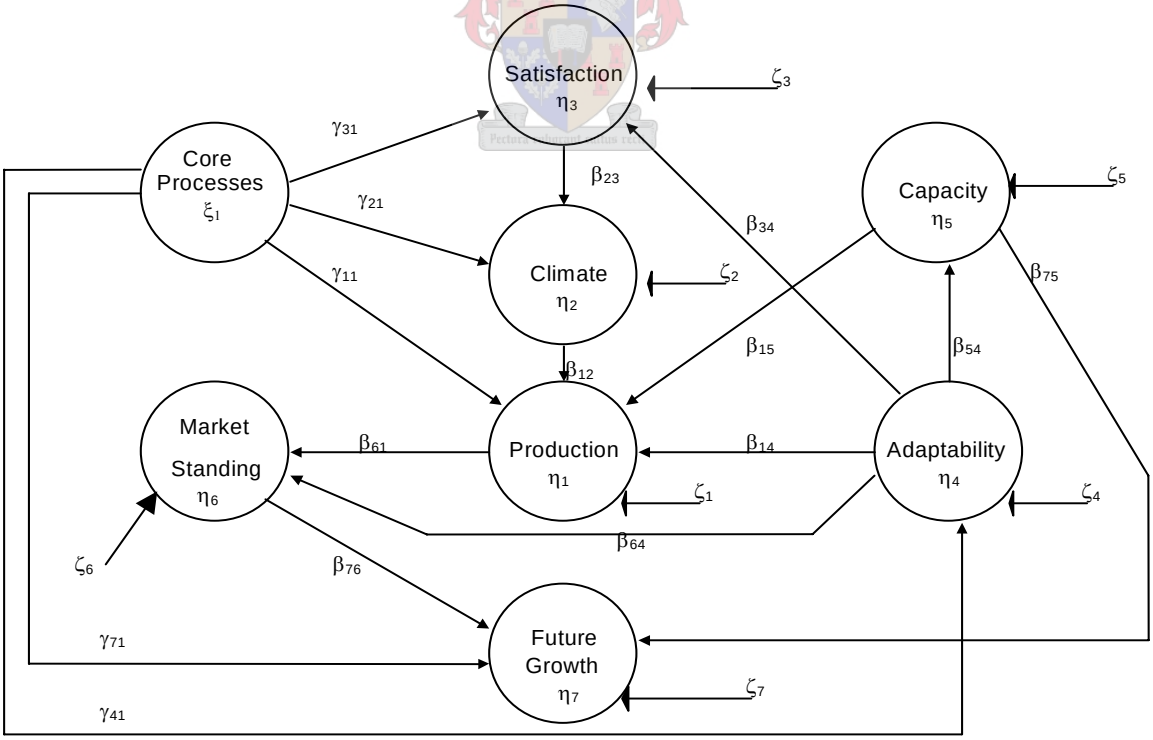


FIGURE 1

THE ORIGINAL PERFORMANCE INDEX STRUCTURAL MODEL

(Theron, Spangenberg & Henning, 2004, p. 37)

However, the Henning *et al.* (2003) study also produced some unexpected findings as the results failed to find support for the hypotheses that there are directional linkages between Capacity and Production & Efficiency, Adaptability and Production & Efficiency. In addition, preliminary analyses by Henning *et al.* (2003) suggested three elaborations to the initially proposed model. In the case of two of the latent variables, factor fission was found to result in conceptually meaningful divisions of the original unit performance dimensions in question. Results of the Henning *et al.* (2003) study suggested the inclusion of an additional path in the original model, representing the influence of market standing on the wealth of resources to which the unit has access. The Henning *et al.* (2003) study also found empirical support for the addition of this path although the *ex post facto* nature of the study's research design precluded the drawing of causal inferences from significant path coefficients.

Based on these findings, Henning *et al.* (2003) proposed a theoretically meaningful refinement of the original PI structural model. The Henning *et al.* (2003) study, however, chose not to follow up on these findings to refine the original unit performance model, but rather first to establish the merits of the simpler, initial model. In a later study, Theron, Spangenberg and Henning (2004) tested the model fit of this elaborated model and found both the original and elaborated PI model to have acceptable model fit. The results of the Theron *et al.* (2004) study mirrored the unexpected findings of Henning *et al.* (2003) as they failed to find support for the hypotheses that there are directional linkages between Climate and Production & Efficiency, as well as Capacity and Production & Efficiency, Adaptability and Production & Efficiency. The results of the previous studies by Henning *et al.* (2003) and Theron *et al.* (2004) highlight a need to further investigate whether the additional alterations to the PI model as proposed by Henning *et al.* (2003) are required. Prior to undertaking further research on the existence of possible interaction effects between the PI latent variables Climate, Adaptability, Capacity and Production and Efficiency it is, however, necessary to cross-validate the measurement model using independent samples within the sample population (Diamantopoulos & Sigauw, 2000). If at least partial measurement invariance would be indicated, the structural invariance of the basic PI model proposed by Henning *et al.* (2003) would moreover have to be examined. Only if the Henning *et al.* (2003) findings that no direct causal linkages exist between Climate and Production & Efficiency as well as between Capacity and Production & Efficiency, and between Adaptability and Production & Efficiency would hold up in a cross validation sample, would further research on the existence of possible interaction effects between the PI latent variables Climate, Adaptability, Capacity and Production & Efficiency be truly justified.

CHAPTER 2
MEASUREMENT INVARIANCE

2.1 RESEARCH OBJECTIVE: ESTABLISHING THE MEASUREMENT INVARIANCE OF THE PERFORMANCE INDEX

Invariance research in general pertains to the question whether measurement and/or structural model parameters differ across different (cultural, gender, racial, age) groups sampled from different populations. Vandenberg (2002, p. 141) illustrates the need for establishing the measurement invariance of an instrument across different populations through the following thought-provoking questions that reflect some of the typical situations researchers may be faced with:

- *Do individuals from different cultures interpret and respond to a given measurement instrument using the same conceptual frame of reference?*
- *Do rating sources (for example in a 360-performance rating situation) use the same definition of performance when rating the same person on the same performance dimension?*
- *Are there individual differences that trigger the use of different frames of references when responding to measures?*
- *Does a process that is purposely altered such as an organisation intervention also change the conceptual frame of reference against which responses are made?*

Given the scenarios alluded to by these questions, it makes sense that establishing the measurement invariance of an instrument across groups should be a prerequisite to conducting substantive cross-group comparisons. Without evidence that supports the invariance of an instrument, the basis for drawing scientific inference should be considered as severely lacking (Horn & McArdle, 1992, cited in Vandenberg & Lance, 2000, p. 9). In addition, if invariance is not yet established for a measure such as the PI or if there is evidence that suggests the measure has significant variance across different groups within the same population, findings of differences between individuals and groups cannot be unambiguously interpreted which in turn raises questions about using the specific instrument within these groups (Steenkamp & Baumgartner, 1998).

Cross-validation is a specific application of the more general form of invariance research (Diamantopoulos & Sigauw, 2000). This study takes the first step in the cross-validation process discussed above, as it poses the research question of whether there is convincing evidence that the current measurement model cross-validates successfully to an independent sample within the same population. In answering this question this study examines if respondents from a different sample from the same target population interpret the PI indicators in a conceptually similar manner, through tests of measurement invariance (Byrne & Watkins, 2003; Diamantopoulos & Siguaw, 2000; Mavondo, Gabbott & Tsarenko, 2003).

Tests of measurement invariance and structural invariance make-up the broader and longer-term process of cross-validation that seeks to establish the invariance of the PI measurement and structural model parameters across independent samples from the target population and, in doing so, support the generalization of the Henning *et al.* (2003) and Theron *et al.* (2004) findings on the PI across different samples from the target population. Tests of measurement invariance test the assumption that the indicator variables are interpreted in a conceptually similar manner by examining the fit of the measurement model across independent samples from the target population. In comparison, tests of structural invariance test the assumption that the underlying theoretical construct elicits the same conceptual frame of reference by examining the fit of the structural model across independent samples of the target population (Byrne & Watkins, 2003; Mavondo *et al.,* 2003; Vandenberg, 2002). As such, establishing the measurement invariance of the PI is a necessary prerequisite to establishing structural invariance (Mavondo *et al.,* 2003; Pousette & Hanse, 2002; Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000).

Although the importance of investigating invariance across qualitatively different groups within the same target population and/or independent samples from the same target population is self-evident, it is not routinely established for measures used in organisational research even though specific aspects of invariance can be established by means of Confirmatory Factor Analysis (CFA) and Structural Equation Modeling (SEM) (Diamamtopoulos & Siguaw, 2000; Vandenberg & Lance, 2000). The general lack of invariance studies is attributed to various factors (Lubke & Muthen, 2004; Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000). Firstly, the array of different types of invariance found in literature and the lack of agreed-upon terminology to refer to these different kinds of invariance is quite bewildering. Secondly, testing for different kinds of invariance often involves considerable methodological complexities including testing measurement models that incorporate the latent and observed variable which researchers tend to

be unfamiliar with. Lastly, there are very few clear guidelines that may be used to ascertain whether or not a measure exhibits adequate invariance.

In recent years some authors, for example Byrne and Watkins (2003); Cheung and Rensvold (2000); Mavondo *et al. (*2003); Steenkamp and Baumgartner (1998); Vandenberg and Lance (2000), and Vandenberg (2002) amongst others, have endeavoured to clarify key invariance issues and proposed best practices for establishing invariance. Although terminology used by these authors continues to differ (especially between researchers focusing on consumer research and those focusing on organisational behaviour) and is likely to confuse readers who are not experts in the field of invariance, there appears to be a narrowing towards a uniform understanding of, and approach towards invariance research.

However, there appears to be an increasing awareness of the need to establish the invariance of instruments used in multigroup contexts or in the same population across time (Steenkamp & Baumgartner, 1998; Vandenberg, 2002; Mavondo *et al.,* 2003). This need is supported by the findings of Vandenberg and Lance (2000) that conducted an extensive review of studies that employed invariance methodology and found that there were many cases in which researchers would have made inaccurate inferences if they had not examined the invariance of the instrument(s) they were using. Studies which examined the differences between groups, measured by differences in group means, were noted by Vandenberg and Lance (2000) as being particularly susceptible to inaccurate conclusions had they not established invariance. Vandenberg (2002) concluded that by establishing the invariance of the instruments being used researchers may conclude with more confidence that the observed differences between groups are a function of the substantive phenomenon being examined and not due to some measurement artefact.

Based on the above discussion, the PI may only be considered invariant across groups if it meets the requirements of both measurement invariance and structural invariance. Although this study only examines the measurement invariance of the PI across independent samples from the target population and not the structural invariance of the PI, it nonetheless forms part of the ongoing research of the leadership-for-performance range of measures designed by Spangenberg and Theron (2004). Thereby this study takes the initial step towards establishing the degree of confidence with which the PI may be used across different groups within the target population.

Similarly, establishing the invariance of the PI will also enhance the confidence of findings of research that links the leadership behaviour to work unit performance.

Furthermore, establishing the invariance (both measurement and structural invariance) of the PI across samples of the same population will justify future research in which the PI may be used for meaningful comparisons between groups, provided the invariance has been established between qualitatively different groups being compared (Durvasula, Andrews, Lysonski, & Netemeyer, 1993; Mavondo *et al.,* 2003). In particular, the theory-based claim that the PI measures work unit performance across all different types of organisations and industries may be examined through future cross-validation studies once invariance of the PI is established within these populations. Other future research may include identifying global attributes of good and poor performing work units over time, or identifying specific changes in performance related to organisational transitions or interventions.

2.2 RESEARCH QUESTIONS: TESTING FOR MEASUREMENT INVARIANCE

Cross-validation of the measurement model refers to an examination of the invariance of the model across two or more random samples from the same population (Mels, 2003) and may be determined by investigating the stability of the model parameter estimates when the model is fitted to two samples from the same population simultaneously (Vandenberg & Lance, 2000). This cross-validation study uses specific measurement invariance tests to answer a sequence of questions or research problems that examine the extent to which the measurement model may be considered measurement invariant or not at all, and to determine the source of variance if it exists (Vandenberg & Lance, 2000). The following series of steps and concomitant research questions capture the essential logic underlying the investigation of measurement invariance. The research questions relevant to this specific study are thereby also explicated.

Step 1: Establish if the measurement model when fitted to each sample independently display reasonable fit when no freed parameters are constrained.

Prior to cross-validating the measurement model it is necessary to first establish whether the model fits on both samples independently. Rejecting the null hypothesis of close fit would indicate that the measurement model does not adequately fit the data of one or both samples, and any further examination of the cross-validation of the PI using these two samples would be

questionable. On the other hand, satisfactory model fit for both samples would justify further cross-validation analysis (Diamantopoulos & Siguaw, 2000). The following research question should thus be answered at the outset:

> *Research question 1: Does the measurement model display acceptable fit on the data of the two samples when fitted in separate, independent confirmatory factor analyses?*

Step 2: Establish if the measurement model, when fitted to the two samples simultaneously in a multi-group analysis with no freed parameters constrained, display reasonable fit.

If the measurement model provides a close fitting account of the process underlying the observed variables the measurement model should show satisfactory fit when fitted to the data of both samples simultaneously with no freed model parameters constrained. Although it is highly unlikely that the model will not show satisfactory fit under these conditions if it was shown to fit both samples independently, results that indicate the contrary would fail to support continuing with the cross-validation study. Demonstrating that the measurement model fits the data of both samples taken from the same population would establish configural invariance (Vandenberg & Lance, 2000). The following research question should thus be posed subsequent to answering the first research question in the affirmative:

> *Research question 2: Does the measurement model display acceptable fit on the data of the two samples when fitted in a single multi-group confirmatory factor analysis without any constraint on parameter equality?*

Step 3: Establish whether the measurement model demonstrated acceptable fit when fitted to the two samples simultaneously in a multi-group analysis with all freed parameters constrained to be equal across the samples.

The most stringent test of measurement invariance tests the null hypothesis (H_{01}: $\Sigma^{g} = \Sigma^{g'}$) that the PI measurement model fits the data the same way across samples from the target population (Diamantopoulos & Siguaw, 2000; Vandenberg & Lance, 2000). The null hypothesis implies that the same underlying process or measurement model is required to explain the observed (in contrast to the reproduced or estimated) population covariance matrices ($\Sigma^{g} = \Sigma^{g'}$) because the observed population covariance matrices are the same. Conversely, if measurement models with different parameter estimates are required to account for the observed covariance in specific

samples it would imply that the covariance matrices differ and that underlying measurement models differ, albeit not to the extent of a lack of configural invariance. If the same measurement model (i.e. configurally the same and in terms of parameter values the same) fits each data set to the same degree of acceptable fit (i.e., close fit) the combined measures of fit would indicate the same degree of acceptable fit. This step tests the null hypothesis that *apriori* pattern of free and fixed factor loadings imposed on the measure's components in terms of the measurement model is equivalent across groups (Horn & McArdle, 1992 cited in Vandenberg & Lance, 2000, p. 12). Failure to reject the null hypothesis would mean the PI may be considered measurement invariant across the samples and subsequent tests of measurement invariance are not required. It is for this reason that this test is termed the omnibus test of measurement invariance.

The omnibus test constitutes a rather severe, stringent test. For most social science research it is highly unlikely that full measurement invariance will be displayed because some difference between the samples is to be expected (Steenkamp & Baumgartner, 1998). As it is almost a forgone conclusion that the null hypothesis will be rejected for this study and given that the results do not provide information on the source of variance, the omnibus test may possibly be considered a somewhat redundant exercise (Vandenberg & Lance, 2000). Despite the odds being against a finding of full measurement invariance it nonetheless constitutes a logical and indispensable part of a systematic and rigorous procedure aimed at investigating measurement invariance. In the context of this study, the omnibus test will thus be conducted in the hope that full measurement invariance might be found but ultimately because it constitutes sound methodological practice.

If the hypothesis of measurement invariance can not be rejected under the configural invariance condition, the model may be said to have cross-validated successfully and further tests of measurement invariance would not be required (Vandenberg & Lance, 2000). This would also imply that that the respondents of each sample employed the same conceptual frame of reference when completing the PI items and provide sufficient evidence of measurement invariance to justify other research that examines group differences in relation to the PI's underlying constructs (Vandenberg & Lance, 2000). The rejection of the null hypothesis would, however, imply that significant difference exist between one or more of the measurement model parameters when the model is fitted to the data of both samples simultaneously. Further measurement invariance tests

would be required to determine the source and extent of this non-equivalence. The following research question is thus indicated:

Research question 3: Does the measurement model display acceptable fit on the data of the two samples when fitted in a single multi-group confirmatory factor analysis and all freed parameter estimates are constrained to be equal?

Step 4: Establish whether the measurement model demonstrated acceptable fit when fitted to the two samples simultaneously in a multi-group analysis with all parameters constrained to be equal across the samples but for the slope of the regression of the indicator variables on the latent variables.

Upon rejection of the full measurement invariance hypothesis the question then needs to be asked whether the non-equivalence exists in the factor loadings of item parcels on latent variables across samples. The null hypothesis states that the factor loadings of item parcels on latent variables are equivalent across both samples (H_{02}: $\Lambda^{g}_{x} = \Lambda^{g}_{x}{}'$). On the one hand, rejection of the null hypothesis would imply that the factor loadings for like items differ across samples, which implies that the content of each item is being perceived and interpreted differently across samples (Byrne & Watkins, 2003). This would constitute a somewhat disappointing outcome of this cross-validation study as the factor loadings really reflect the core of the measurement process. Logically the items would be expected to operate in the same manner across independent random samples from the target population (Pousette & Hanse, 2002; Vandenberg & Lance, 2000). It would, however, not be an altogether improbable outcome as H_{02} would only be tested if H_{01} would have been rejected and thus significant differences in some model parameters have to exist. Rejection of H_{02} due to a limited number of significant differences in factor loadings would indicate partial metric invariance. On the other hand, failure to reject the null hypothesis that the factor loadings are equal across both samples (H_{02}) would mean that the measurement model displays metric invariance. This would be a fairly satisfactory outcome as it would support the conclusion that the item parcels operate in the same way across samples in the way they reflect the underlying latent variables they are meant to reflect. In addition, this outcome would justify further research that examines group differences in relation to the PI's underlying constructs for similar samples within the target population. At least partial metric invariance of the PI would indicate that the PI measurement model displays sufficient measurement invariance within the target population to warrant further examination of the structural relationship between the latent dimensions, including tests of structural invariance (Byrne & Watkins, 2003; Diamamtopoulos & Siguaw, 2000; Mavondo *et al.,* 2003). In doing so the differences in factor

loadings would, however, have to be taken into account. The following research question is thus indicated:

Research question 4: Are the factor loadings of item parcels invariant across the samples?

Failure to reject the H_{02} metric equivalence null hypothesis would indicate that significant differences in parameter estimates that were detected by previous measurement invariance tests do not exist within the factor loadings. The source and strength of these differences would thus still need to be determined as they have to exist elsewhere in the measurement model. Additional tests of measurement invariance are therefore required to examine the differences in parameter estimates of the model's factor covariances and the model's measurement error variances when fitted to both samples simultaneously (Vandenberg & Lance, 2000).

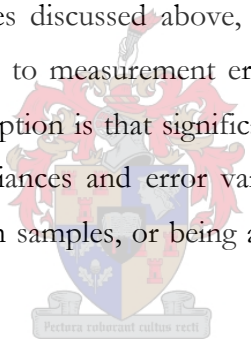
Step 5: Establish whether the lifting of the equality constraint on the factor covariances and variances significantly improves the fit of the measurement model when fitted to the two samples simultaneously in a multi-group analysis.

Testing for the equivalence of factor covariances between groups tests the null hypothesis that the phi matrices are invariant across both samples (H_{03}: $\Phi^{g}_{ij} = \Phi^{g}_{ij}$). Failure to reject the H_{03} null hypothesis would imply that both samples use "equivalent ranges of the construct continuum to respond to the indicators reflecting the construct" (Vandenberg & Lance, 2000, p. 39). This would add credence to the finding of at least partial metric invariance because it would imply that the variance in the measurement model might be attributed to variance in measurement error. On the other hand, rejection of the H_{03} null hypothesis would indicate that significant variance exists between the factor covariances across samples. This outcome is not desirable as it would serve to somewhat devalue the conclusion of at least partial metric invariance. The following research question is thus indicated:

Research question 5: Can significant differences between samples be attributed to differences in factor covariances between, and variances of, latent variables across samples?

Step 6: Establish whether the lifting of the equality constraint on the measurement error variances significantly improves the fit of the measurement model when fitted to the two samples simultaneously in a multi-group analysis.

In comparison, it would be far more desirable to be able to attribute the source of significant variance between the samples to error variances. This may be established by testing the null hypothesis of equal variance in the error terms associated with the indicator variables across groups (H_{04}: $\theta_{\delta}{}^{g}_{j} = \theta_{\delta}{}^{g}_{j}{}'$). Rejection or acceptance of the null hypothesis would need to be interpreted in relation to the difference in factor covariances. Failure to reject the null hypothesis for both tests of equal error variances and equal factor covariances would provide evidence that both samples respond to the indicator variables in an equivalent manner, in that the no significant variance exists across samples in terms of the error terms or factor covariances associated with the indicator variables. This would be the most desirable outcome as it would suggest the operation of the measurement model does not differ greatly across both samples, thus supporting the conclusion that the measurement model is sufficiently invariant across the samples. If no significant difference was found to exist in the factor covariances then all of the variance in the measurement model fit between the two samples may be attributed to non-equivalent error variances across samples. This would be a better outcome than having to reject the null hypothesis of equal factor covariances discussed above, as it is more desirable to be able to attribute differences between samples to measurement error rather than to differences in item response across samples. A further option is that significant differences across samples may be found to exist for both factor covariances and error variances, again not as desirable as not finding significant differences between samples, or being able to attribute significant differences to measurement error.



Research question 6: Can significant differences between samples be attributed to variance in the error variances across samples, or to both error variances and factor covariances across samples?

The foregoing proposed procedure consistently uses the fully unconstrained model as the baseline model in the multiple group analyses used to determine whether measurement invariance exists, and if not in which facet/facets of the measurement model the differences reside. The fully or partially constrained measurement models are therefore compared each time to the same fully unconstrained measurement model to determine whether the full or partial equality constraints result in a significant deterioration in fit. The question, however, needs to be considered whether a moving baseline model should not be used when the measurement invariance null hypothesis is rejected to determine how the measurement model parameters differ across the samples? This study would justify the use of a fixed baseline model by arguing that reality expresses itself in the fully unconstrained model. If, for example, the factor loadings of

item parcels on latent variables would not differ across samples, a model in which factor loadings are constrained to be equal across samples will not fit significantly poorer than the unconstrained model. Moreover, in subsequent analyses aimed at locating the source of measurement invariance there would be no need to compare a model in which both the lambda-X and phi matrices are constrained to be equal to a model in which only lambda-X is constrained. The lambda-X equality constraint is naturally built into the fully unconstrained model.

The converse also could be argued. If, for example, the factor loadings of item parcels on latent variables would differ significantly across samples, a model in which factor loadings are constrained to be equal across samples will fit significantly poorer than the fully unconstrained model. Moreover, in subsequent analyses aimed at locating further sources of measurement invariance there would be no need to compare a model in which the lambda-X matrix is unconstrained but the phi matrix is constrained to be equal to a fully unconstrained model. The lambda-X inequality is naturally built into the fully unconstrained model.

2.3 STATISTICAL ANALYSIS TECHNIQUE

Structural Equation Modelling (SEM) is used to perform a series of confirmatory factor analyses on the subscales of the PI using LISREL 8.53 for Windows (Du Toit & Du Toit, 2001; Jöreskog & Sörbom, 1998). As stated by Steenkamp and Baumgartner (1998) there is general consensus that LISREL's multigroup confirmatory factor analysis model represents that most powerful and versatile approach to testing for multiple sampling applications of measurement invariance.

As an analysis technique SEM also has certain advantages that apply to this research Kelloway (1998). Firstly, SEM affords social science researchers the opportunity to determine how well measures, used to represent latent constructs, reflect the intended constructs in a more rigorous and parsimonious way than the techniques of exploratory factor analysis traditionally employed by enabling researchers to specify structural relationships among the indicator variables and the specific latent variables they are meant to reflect (Bollen & Long, 1993; Kelloway, 1998). SEM allows for explicit tests of hypothesis relating to the overall quality of the factor solutions, as well as the specific parameters comprising the model. Secondly, SEM assists researchers in the use of complex predictive models by allowing for the testing and specification of these more complex "path" models as an entity in addition to testing the components comprising the model. Lastly, SEM provides for the estimation of the strength of the relationship that exists

between latent variables, without being moderated by measurement error (Bollen & Long, 1993). As such, SEM may be considered a flexible, yet powerful approach to investigating various forms of measurement invariance in first- and higher-order measurement models.

CHAPTER 3

RESEARCH METHODOLOGY AND PREPARATORY DATA ANALYSES

3.1 SAMPLING STRATEGY

Two independent samples of completed PI questionnaires were required for this study. These were collected through non-probability sampling procedures. To be included in this research, unit managers had to manage work units that met the requirements of a work unit as defined in the introduction to the paper, and were in their current position for at least six months. As the PI is a 360° instrument, work units were rated by the unit leader, as well as their superiors, peers and subordinates. However, the need for as large a sample size as possible, necessitated a deviation from this ideal in some of cases, although this deviation was considered to be acceptable practice because the research requires the analysis of data on an individual level, and not on a collective work unit level.

3.1.1 Sample A

Sample A combined two data sets from previous research and includes a total of 313 completed PI questionnaires. Of these completed questionnaires, 256 were gathered from part-time MBA students of the Graduate School of Business at the University of Stellenbosch during the 1998, 1999 and 2000 intakes. These MBA students occupied full-time positions in middle or senior management. Out of a possible number of 115 eligible work unit managers, 60 participated in the study, which represents a satisfactory 52% participation (Spangenberg & Theron, 2002). The other 47 completed questionnaires came from three different functional departments in a large fast moving consumable goods (FMCG) company and represented 47% participation as 100 questionnaires were sent out (Henning *et al.*,, 2003). No information on the number of completed units was available for the 47 completed questionnaires. Although no demographic information pertaining to Sample A was available, it may be assumed that the sample is a fairly good representation of the target population because the MBA students are likely to represent diverse professions across different companies and industries in South Africa.

3.1.2 Sample B

Sample B included 393 completed PI questionnaires rating the performance of 65 work units and was obtained through a Management Development Programme at the University of Stellenbosch which included a PI evaluation. Out of 86 course delegates, 65 (7 female: 9%, 58 male: 91%) met the requirements to participate in the study. These delegates occupied full-time positions in middle management and senior management within a large multinational mining group, and represented various professions within the mining industry such as engineering, finances, purchasing, logistics, safety, and human resources. Delegates represented a wide array of ethnic groups and nationalities, and work units were spread across six countries as indicated in Table 2 which provides further qualitative information for Sample B. As the PI evaluation formed part of their development programme, delegates were motivated to participate. A total of 556 questionnaires were sent to unit managers and respondents, and 393 completed questionnaires were returned. This figure represents a 71% response rate that can be considered quite satisfactory.

TABLE 2

QUALITATIVE INFORMATION FOR SAMPLE B

	No.	%
Respondents at the various levels:		
Unit managers	52	13 %
Superiors	68	17 %
Peers	116	30 %
Followers	157	40 %
Total respondents	*393*	
Unit managers' position:		
Middle management	40	77 %
Senior management	12	23 %
Total	*52*	
Gender of unit managers rated:		
Male	58	91 %
Female	7	9 %
Total no. of work unit rated	*65*	
Location of work unit operations:		
England	2	3 %
Botswana	2	3 %
Australia	4	6 %
South Africa	44	67 %
Ireland	5	7 %
Namibia	9	14 %
Total number of work units rated	*65*	

3.1.3 Possible limitations of sampling method

To convincingly demonstrate that the PI functions effectively within the target population would require two independent random samples taken from the same population rather than non-probability samples. This ideal was clearly not met as both samples may not be considered truly representative of the target population of middle and senior managers within multiple companies and industries. For example, Sample B only includes information from one company and one industry which is likely to have idiosyncratic cultures and operating procedures that may influence the manner in which the respondents perceive and rate work units. As such, Sample B would be said to better represent the target population if more industries and companies within these industries were included, a consideration for future studies. It would also better represent the target population if it included more female unit managers than the 9% included in the current sample. In comparison, Sample A includes participants from various companies and industries and therefore may be considered to be more representative of the target population.

Based on the above, both samples cannot be said to constitute a representative section of the population of work units, which precludes the possibility of reaching any definitive conclusion on whether the PI could be used for all organisations and industries across the target population. Nonetheless sufficient fit between the measurement models of both data sets would constitute relevant, albeit limited evidence that the PI meets the requirements of measurement invariance and hence investigation into the PI's structural invariance would be warranted. Sufficient fit between the measurement models also would provide limited evidence that the PI may be used to assess the perceived performance or work units in the target population.

The fact that the two samples do not constitute two independent probability samples from the same population moreover presents the PI measurement model with a more severe cross-validation challenge than it would have faced had the samples been truly independent random samples from the same population. Instead of cross-validating the measurement model across samples A and B a better option probably would have been to combine the data from the two samples and to randomly create two samples from the combined data set for the purpose of cross-validation. Although the resultant samples still would not constitute representative samples from the target population for which the PI had been developed, the cross-validation would at least provide a more valid indication of the invariance of the measurement model parameters.

3.2 MISSING VALUES

The most suitable method for managing missing values had to be chosen prior to data analysis. Table 3 summarises the missing values per dimension and Table 4 (page 29) presents the number of missing values per item. Sample B had fewer (8,84%) missing values than Sample A (9,96%). A possible explanation for this improvement is that the data for Sample B was collected using a computerised version of the PI which does not allow respondents the option to leave a question blank although they may choose a non-observable option. Unfortunately the paper-and-pencil version that was used to collect data for Sample A cannot impose a forced choice on respondents as the computerised version is able to.

TABLE 3

SUMMARY OF MISSING VALUES PER DIMENSION

	% missing values		
Dimension	**Sample A**	**Sample B**	**Sample A & B combined**
N	313	393	706
Product	4,28%	2,14%	3,09%
Core people	2,52%	1,84%	2,14%
Climate	1,60%	0,55%	1,01%
Satisfaction	2,88%	2,88%	2,88%
Adaptability	4,38%	1,85%	2,97%
Capacity	10,68%	10,58%	10,62%
Market standing	21,68%	27,41%	24,87%
Future growth	37,12%	33,79%	35,92%
Total number of missing values	**1640**	**1946**	**3586**
Missing values as a % of total values	**9,36%**	**8,84%**	**9,07%**

Selecting the most suitable method of managing missing values was not a simple task as different methods require certain assumptions about the nature of the data and the reasons for the missing values that are not openly acknowledged or observable during the data gathering phase (Pigott, 2001). This section discusses the advantages and disadvantages of various methods for managing missing values in relation to this research and why imputation by matching was selected as the most appropriate response in this case. In particular two questions had to be considered, firstly whether the reasons why PI values are missing can be ignored which refers to the assumption of

an ignorable response mechanism, and secondly whether the distribution of indicator variables may be assumed to be multivariate normal.

3.2.1 The assumption of an ignorable response mechanism (MAR/MCAR)

In their seminal work on the analysis on incomplete data, Little and Rubin (1987, cited in Davey, Shanahan & Schafer, 2001) distinguish between two types of ignorable missing data, namely missing completely at random (MCAR) and missing at random (MAR). MAR is a related but weaker assumption than MCAR which holds a conditional assumption that the missing values have the same distribution as the non-missing values of the observed sample (Enders & Bandalos, 2001; Pigott, 2001). If missing values are MCAR, cases with missing values are indistinguishable from cases with complete data. In contrast, if missing values are MAR, cases with missing values differ from cases with complete data. Whether a case returns a missing value on one or more variables is, however, not related to the state of the variables in question but rather to one or more other variables in the data set (Information Technology Services, 2005). If missing values are considered to be MAR the reasons for the missing values may be ignored during the analysis of the data, which in turn allows the researcher to use most model-based methods of imputing missing values (Pigott, 2001; Roth, 1994). The pattern of missing values across cases would be considered non-ignorable if that pattern is only explainable in terms of the state of the variables on which missing values are returned (Information Technology Services, 2005).

It is however difficult to obtain empirical evidence about whether the data in this research is MCAR or even MAR. Without empirical evidence possible reasons for missing data have to be inferred in order to take the best decision (Pigott, 2001). In this research, a number of reasons could be inferred that suggested it would not be appropriate to consider the missing values for the PI to be missing completely at random (MCAR) or missing at random (MAR), and that the reasons for the missing data should therefore be considered non-ignorable (Little & Rubin, 1987, cited in Davey *et al.,* 2001). On the one hand the PI provides respondents with the option to select a non-observable response for an item either if they feel the item is relevant to the work unit but they are not in a position to rate the work unit on this item, or if they perceive the item to be irrelevant or non-applicable to the work unit.

On the other hand though, there may well be additional reasons why respondents selected the non-observable option which directly relates to the value of that variable being measured. For example, if a Sample for company X has many missing values for Projected Growth (dimension 8) across various work units it may indicate the company as a whole is experiencing declining growth. Pigott (2001) suggests that for cases in which the missing value is directly related to the value of that variable, the missing values should be considered not to be MAR and therefore should be non-ignorable. Furthermore, Switzer, Roth and Switzer (1998) conclude there is reason to believe that non-random data loss might decrease covariance and power more significantly than in the case of random data loss.

In contrast, research by Davey, *et al.* (2001) illustrated that any attempt to identify and correct for selective non-response will represent an improvement in the accuracy of results over making no attempt at all. They concluded that even when data are not strictly MAR, it is believed this assumption is likely to represent a reasonable approximation and therefore may be assumed. Similarly, Roth (1994) found little differences in the parameter estimates and the answers to research questions when less than 10% of the data was missing in random or systematic patterns.

In summary, even though it may be argued that the data in this research is not MAR, Table 3 illustrates that the missing values as a percentage of total values for Sample A is 9,36%, for Sample B is 8,84%, and for the Combined Sample is 9,07%, all of which are below 10%. Therefore all methods of managing missing values, including those which require an assumption of MAR, were considered in this study. These methods were categorised as either: deletion methods, model based (distributional) methods, and non-model based methods of imputing missing values.

3.2.2 Deletion methods

List-wise and pair-wise deletion

After a review of current literature, list-wise deletion (complete-case analysis) and pair-wise deletion (available-case analysis) were not perceived to be the best options for managing the missing values problem. List-wise deletion involves deleting complete cases where there is missing values for any of the variables, whilst pair-wise deletion involves deleting cases only for analysis on variables where values are missing. As such, both list-wise and pair-wise deletion methods result in a large loss of data in order for enough complete cases to remain in order to

estimate the desired model. For example, in this study list-wise deletion would have resulted in Sample A being reduced to 81, and Sample B being reduced to 151 cases. Pair-wise deletion would have resulted in covariance matrices with extreme variation in N-values. Subscales would have been reduced to a minimum of 77 and maximum of 251 values for Sample A, and similarly a minimum of 173 and a maximum of 366 values for Sample B.

Using only completed data for list-wise deletion, any partial information from incomplete cases that may be quite valuable is ignored. This may distort the representivity of the original sample, especially if the subjects who are included in the analysis are systematically different from those who were excluded in terms of one or more key variables (Raghunathan, 2004). Furthermore, pair-wise deletion has been found to potentially produce invalid estimates due to the varying samples used to estimate parameters (Pigott, 2001). In particular, pair-wise deletion may yield non-positive definite correlation and covariance matrices in situations where three or more variables are involved. Covariance or correlation matrices that are not positive definite may cause discrepancy function values to become negative which violates the requirement for discrepancy functions to be bounded below by zero (Browne, 1982 cited in Kaplan, 2000 pp. 88-89).

Pair-wise deletion based correlation matrices also do not maximise any proper likelihood function and with respect to structural equation modelling, this will probably affect the chi-square goodness-of-fit test (Kaplan, 2000). Even in the case where pair-wise deletion does not result in non-positive definite matrices, MCAR is assumed to hold for the subset of observations that remain (Kaplan, 2000). List-wise and pair-wise deletion are therefore also likely to yield biased estimates in situations and are not recommended unless the amount of missing data is small and MAR can be assumed to hold (Pigott, 2001). Given the above information, a more suitable method for managing missing data was sought for this research.

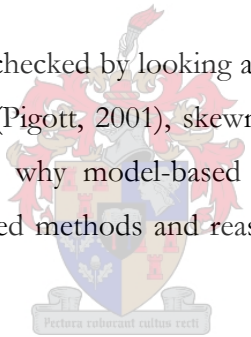
3.2.3 Model based (distributional) methods

Recent research strongly suggests that model based methods, in particular Full Information Maximum Likelihood and Multiple Imputations, are progressive techniques that have advantages over the more traditionally used deletion and non-model based methods (Enders & Bandalos, 2001; Pigott, 2001; Sartori, Salvan & Thomaseth, 2005; Schafer, 1999). Before these methods may be employed the research data has to meet the assumptions of multivariate normality and MAR discussed above.

The assumption of multivariate normality

To assume data has a multivariate normal distribution requires the data to be continuous and therefore to preclude the use of nominal (non-ordered categorical) and ordinal variables. This is because these methods use the multivariate relationship between variables to compute estimates for missing data (Pigott, 2001). As the individual item data in this study are ordinal and not continuous they do not meet this assumption. However, as Schafer (1999) found through comparative research on different methods of managing missing data, this assumption can be relaxed to the assumption that the data is multivariate normal on condition that the categorical variables in the model are completely observed. If categorical variables in the data have high rates of missing observations then methods using the multivariate normal assumption should not be used (Pigott, 2001). In contrast, if only small amounts of missing observations or values are completely observed, model-based methods seems to be fairly robust in situations whereby data has a moderate departure from normality (Schafer, 1999).

Whilst multivariate normality may be checked by looking at the skewness of the data per variable using techniques such as histograms (Pigott, 2001), skewness was not analysed as the literature reviewed revealed additional reasons why model-based methods are not well-suited to this particular research. These model based methods and reasons for not using them are discussed below.



Full information maximum likelihood

Full Information Maximum Likelihood (FIML) uses an iterative solution, termed the EM algorithm, to compute a case-wise likelihood function using only those variables that are observed for specified cases. By doing this it obtains estimates of missing values based on the incomplete observed data to maximise the observed data likelihood (Enders & Bandalos, 2001; Raghunathan, 2004). Research indicates that FIML is a progressive technique with advantages over other methods (Enders & Bandalos, 2001; Roth, 1994). For example, it has been found to reduce the bias that would result from the list-wise or pair-wise deletion of cases (Enders & Bandalos, 2001), and its estimation procedure is viewed as more efficient than that of other imputation methods (Du Toit & Mels, 2002). In addition, Kaplan (1995) examined Chi-squared test for the related multiple–group approach and found that the mean, variance and rejection rules of the empirical chi-square distribution closely matched those of the appropriate central chi-square distribution.

At the same time, researchers cite problems with using FIML such as the computation difficulties (Pigott, 2001) and the possibility of model misfit problems if the distribution is non-normal (Arbuckle, 1996 cited in Enders & Bandalos, 2001, p.435). Moreover, a full sample is a necessary requirement of this research for the necessary analysis of the PI model and FIML only calculates the expected values of sufficient statistics and does not impute missing values. Using FIML would thereby limit the analyses that may be conducted when compared with a complete data set, and it was not deemed suitable for this research.

Multiple imputation

In comparison to FIML, completed data sets are possible through multiple imputation (MI). Each of the multiple imputations produces a completed data set, which has to be analysed separately in order to obtain multiple estimates of the parameters of the model (Davey *et al.,* 2001; Raghunathan, 2004; Schafer, 1999). The main advantage of MI as stated by Raghunathan (2004) is that it reflects the uncertainty in the estimates, whilst delivering plausible values. In other words it corrects for bias by conducting several imputations for each missing value (Sartori *et al.,* 2005).

Although MI is therefore considered quite a robust method (Schafer, 1999; Schafer & Olsen, 1998), the model used to generate the imputations will however only ever be approximately true. Additional shortcomings of MI are firstly, that it involves multiple and complex statistical analysis which means that it is often an impractical and cumbersome method to use for research. Typically up to ten imputed data bases are created. Performing the multi-group measurement invariance analyses up to ten times and combining the findings into a summary finding seems somewhat unrealistic. Secondly MI procedures available in LISREL 8.54 assume that the values are MAR and that observed variables are continuous and follow a multivariate normal distribution (Du Toit & Du Toit, 2001). As discussed above individual responses to the PI items are given on a five point Likert scale and therefore should be viewed as ordinal in nature (Jöreskog & Sörbom, 1996a).

In sum, model-based methods are widely considered more appropriate to use than non-model based methods and deletion methods. Nonetheless, the data in this research does not meet the prerequisite of MAR and multivariate normality necessary for model-based methods, and therefore alternative non-model based methods of imputing missing values were explored.

3.2.4 Non-model based methods imputing of missing values

Non-model based methods of imputing missing values include single mean imputation and imputation by matching. When compared with the model-based methods, non-model based methods are attractive options to use in this research as the data does not have to meet the assumptions of multivariate normality and categorical variables may be used. Likewise, non-model based methods save a large amount of data that would be lost by employing a deletion method. However, they are not without their limitations.

Single mean imputation

Single mean imputation involves a straightforward procedure of replacing all missing values on a variable with the mean of all cases on that variable. This method was not considered suitable as it will change the distribution of each variable with missing values by decreasing the variance. In other words, the variance of these variables will be underestimated, and as Theron and Spangenberg (2004, p. 23) describe, "effectively wash out most of the structure that exists in the data". As such, single mean imputation is considered to be a fairly crude method and is not recommended as it will always produce biased results (Pigott, 2001).

Imputation by matching (Similar response pattern imputation)

Imputation by matching, often referred to as similar response pattern imputation, attempts to impute values from another case with similar observed values. This is achieved by using a minimization criterion on a set of matching variables (Jöreskog & Sörbom, 1993). If no observation exists that has complete data on the set of matching variables, imputation does not take place for that case (Enders & Bandalos, 2001). Imputation by matching is not without its limitations as it is possible that unless the data are MCAR imputation by matching data, sets will be biased. However, the estimated data has the benefit of preserving deviations from the mean and the shape of the distribution (Little, Cunningham, Shahar & Widaman, 2002) and therefore will not attenuate correlations as much as mean substitutions (Roth, 1994).

The most suitable method for managing missing values for this research is the one that may be considered to enhance the inferential validity of research results the most (Raghunathan, 2004). Given the above comparison of all the methods, imputation by matching was deemed the most suitable method, particularly because missing values on the PI questionnaires do not necessarily

meet the assumption of ignorable response mechanism or the assumption of multivariate normality as discussed previously.

Although the ideal is to use matching variables that will not be used in the confirmatory factor analysis, this was not possible in this study. Rather, the items least plagued by missing values were identified. Table 4 presents the number of missing values per item with items that were identified as matching variables for either Sample A or for Sample B highlighted in bold font.

TABLE 4

NUMBER OF MISSING VALUES PER ITEM

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7
Sample A	14	10	8	28	7	6	**3**
Sample B	**0**	**1**	4	36	**1**	4	**1**
Total A&B	14	11	12	64	8	10	4
	Item 8	Item 9	Item 10	Item 11	Item 12	Item 13	Item 14
Sample A	**4**	**4**	8	8	**1**	**4**	33
Sample B	**1**	5	2	5	9	6	32
Total A&B	5	9	10	13	10	10	65
	Item 15	Item 16	Item 17	Item 18	Item 19	Item 20	Item 21
Sample A	**0**	7	5	**4**	7	5	7
Sample B	**1**	**0**	**0**	6	6	**1**	**1**
Total A&B	1	7	5	10	13	6	8
	Item 22	Item 23	Item 24	Item 25	Item 26	Item 27	Item 28
Sample A	5	11	20	15	9	5	6
Sample B	7	13	26	20	7	5	6
Total A&B	12	24	46	35	16	10	12
	Item 29	Item 30	Item 31	Item 32	Item 33	Item 34	Item 35
Sample A	5	5	**4**	13	14	17	22
Sample B	9	9	**1**	6	7	4	20
Total A&B	14	14	5	19	21	21	42
	Item 36	Item 37	Item 38	Item 39	Item 40	Item 41	Item 42
Sample A	10	16	18	103	14	16	11
Sample B	5	8	29	120	10	21	15
Total A&B	15	24	47	223	24	37	26
	Item 43	Item 44	Item 45	Item 46	Item 47	Item 48	Item 49
Sample A	28	44	124	87	96	59	58
Sample B	26	70	175	157	171	84	85
Total A&B	54	114	299	244	267	143	143
	Item 50	Item 51	Item 52	Item 53	Item 54	Item 55	Item 56
Sample A	28	23	163	155	116	78	69
Sample B	44	38	168	192	111	71	84
Total A&B	72	61	331	347	227	149	153

Missing values were imputed using the PRELIS programme (Jöreskog & Sörbom, 1996b). To impute missing values for Sample A, a set of 8 items with four or less missing values per item were defined to serve as matching variables. After imputation 277 cases with observations on all 56 items remained in Sample A and 36 cases were eliminated. To impute missing values for Sample B a set of 11 items with one or no missing values per item were defined to serve as matching variables. Only 18 cases did not have values for all 56 items after imputation and had to be eliminated. Sample B therefore included 375 complete cases. Substantially more cases were retained for both Sample A and Sample B using imputation by matching than would have been retained through either listwise or pairwise deletion methods.

3.3 ITEM ANALYSIS

The architecture of the PI reflects the intention to construct essentially one-dimensional sets of items to reflect variance in each of the eight latent variables that collectively constitute the domain of work unit performance. This intention should be verified through item and uni-dimensionality analysis. Item analysis is mostly used to identify and eliminate items from an instrument that do not contribute to an internally consistent description of the sub-scale in question. The selection, substitution, or revision of items identified by item analysis assists test developers to improve instruments' validity and reliability (Anastasi & Urbina, 1997). In this case the item and dimensionality analyses were primarily performed to screen individual items prior to their inclusion in item parcels representing the latent variables of interest as aggregate indicator variables.

3.3.1 Item statistics

Item analysis was conducted on each sample before and after imputation. Each of the eight PI sub-scales were item analysed independently through the SPSS Reliability Procedure (SPSS 13 for Windows, 2005) to identify and eliminate items not contributing to an internally consistent description of the unit performance facet in question. Two items were flagged as potentially problematic for Sample A. Item #24 refers to the *level of satisfaction employees show with regards to salary and fringe benefits*, whilst item #41 refers to *the effectiveness of diversity policies or programmes towards ensuring continuous development of all staff, including the empowerment of previously disadvantaged people.* However, the corrected item total correlations (0,466 and 0,396) and the increase in alpha affected by the removal of the items (0,04 and 0,01) argued against the removal of these items.

Likewise, two items were flagged as potentially problematic for Sample B, including item #24 that was also identified as problematic for Sample A, and item #39 which referred to the *growth or decline of profits over the last five year period.* The corrected item total correlations (0,441 and 0,286) and the increase in alpha affected by the removal of the items (0,05 and 0,09), however, did not support the removal of these items.

3.3.2 Sub-scale reliability

A summary of results of the item analyses for Sample A and Sample B are shown in Table 5 and Table 6 respectively. For Sample A, five of the eight sub-scales returned Cronbach alpha values greater than 0,80, with Future Growth falling marginally below this cut off value (0,792). Capacity and Product displayed somewhat lower item homogeneity with Cronbach alpha values of 0,724 and 0,732. In general the homogeneity found for all subscales on Sample A may be considered relatively high and therefore reasonably acceptable.

In a similar vein, Sample B showed a relatively high level of homogeneity as five of the eight sub-scales returned Cronbach alpha values greater than 0,80. Market Standing fell slightly below this cut off value (0,794). However, the remaining two subscales Production and Efficiency (0,777) and Future Growth (0,748) scales again provided some reason for concern.

In these cases the effect that imputation has on reliability coefficients had to be considered. As Table 5 and Table 6 highlight, imputation has an attenuating affect on internal consistency calculations when the number of valid cases increased with imputation, and the opposite affect when the number of valid cases decreased with imputation. Thus, imputation served to improve the reliability coefficient of the Production and Efficiency sub-scale, but decreased the reliability coefficient of Future Growth. In general, though, given the intended use of the PI as a comprehensive criterion measure against which to validate leadership and other competency assessments, and given the number of items included in each sub-scale, the relatively high internal consistency found for most sub-scales both before and after imputation, is considered reasonably satisfactory.

TABLE 5

RELIABILITY OF PI SUB-SCALES FOR SAMPLE A

		Before Imputation				After imputation (N=277)		
Scale	Number of items	Valid cases	Alpha	Mean	Variance	Alpha	Mean	Variance
Product	5	304	0,732	19,07	10,441	0,777	18,8195	8,721
Core people	9	308	0,833	31,77	37,857	0,852	31,3538	35,012
Climate	7	305	0,866	25,47	25,895	0,884	25,2635	26,064
Satisfaction	9	307	0,868	31,54	42,243	0,892	31,0108	38,388
Adaptability	7	301	0,811	24,92	24,110	0,824	24,3069	21,018
Capacity	7	291	0,724	24,41	27,974	0,815	22,8267	21,999
Market standing	7	265	0,830	27,23	39,373	0,838	24,8989	21,743
Future growth	5	258	0,792	20,29	31,329	0,748	17,0072	10,370

TABLE 6

RELIABILITY OF PI SUB-SCALES FOR SAMPLE B

		Before Imputation				After imputation (N=375)		
Scale	Number of items	Valid cases	Alpha	Mean	Variance	Alpha	Mean	Variance
Product	5	353	0,805	19,01	8,395	0,803	18,9787	8,363
Core people	9	351	0,855	31,79	39,632	0,845	31,8747	29,003
Climate	7	384	0,892	25,07	23,068	0,890	25,0373	23,020
Satisfaction	9	358	0,894	30,17	34,900	0,891	30,2213	34,772
Adaptability	7	360	0,848	23,87	18,851	0,839	23,7920	18,449
Capacity	7	250	0,779	23,34	16,571	0,758	23,2640	15,486
Market standing	7	191	0,811	23,91	17,966	0,794	24,1093	17,082
Future growth	5	191	0,799	16,92	9,867	0,745	16,8160	8,578

3.4 DIMENSIONALITY ANALYSIS

Conducting both item- and dimensionality analyses are important prerequisites for ensuring valid and justifiable conclusions of this study, particularly because the measurement model implied by the PI will not be tested by operationalising the eight unit performance latent dimensions in terms of the individual PI items but rather in terms of item parcels. The PI was developed with the objective to construct uni-dimensional sets of items to reflect variance in each of the eight latent variables collectively comprising the work unit performance domain. As such, dimensionality analysis serves to confirm the uni-dimensionality of each of the PI sub-scales. The relatively favourable item analysis statistics reported above (especially the Cronbach alpha values) provide insufficient evidence to conclude that the intention to construct uni-dimensional sets of items to reflect variance in each of the eight latent variables had been successful. Sub-scales that fail the test of uni-dimensionality have to be analysed to determine if specific items with inadequate factor loadings should be removed and the dimensionality analysis repeated, or if

heterogeneous sub-scales have to be split into two or more homogeneous subsets of items thereby forcing the measurement and structural models to be revised (Anastasi & Urbina, 1997).

To confirm the uni-dimensionality of each sub-scale unrestricted principal axis factor analysis with Varimax rotation was performed on each of the eight PI sub-scales individually for each sample. Principal axis factor analysis was chosen over principal components analysis as the statistical calculations in the former allows for the presence of measurement error, an intrinsic aspect of research into human behaviour (Stewart, 2001). In contrast though, Varimax rotation was chosen over oblique rotation even though oblique rotation is considered a theoretically superior method to orthogonal rotation techniques as it had been found to provide better fit when interrelations between variables being measured are expected (Kerlinger & Lee, 2000; Stewart, 2001). Varimax rotation was selected because the interpretation of oblique rotation is complex (Tabachnick & Fidell, 1989). SPSS 13 for Windows (2005) was used for these analyses and the eigenvalue-greater-than-unity rule of thumb was used to determine the number of factors to extract.

In addition, there is a possibility that only artefact factors which reflect differences in item difficulty value or variance may be extracted when uni-dimensionality is examined by performing factor analysis on a matrix of product moment correlations (Hulin, Drasgow & Parsons, 1983). Descriptive statistics were, therefore, calculated for the items of each sub-scale in order to determine the possibility of multiple factors appearing as an artefact of differential item characteristics such as skewness.

3.4.1 Item factor loadings for Sample A

Item factor loadings were generally satisfactory for Sample A as they varied between 0,422 and 0,820, with 95% percent (53 items) exceeding the 0,500 cut off point. The three items that returned a factor loading below 0,500 were items #41 (0,422), #55 (0,445) and #56 (0,470). Items #55 and #56 contribute towards the Future Growth sub-scale along with items #52, #53 and #54. A possible reason for low factor loadings may be that the number of missing values for items #52, # 53 and #54 were very high (52%; 49%; 37%). A similar pattern was found for Sample B whereby items #52, #53 and #54 had very high percentages of missing items (43%; 49%; 28%) and items #55 and #56 also displayed low factor loadings of 0,491 and 0,519, respectively. As mentioned above, the imputation of missing values appears to affect certain statistics such as the coefficient of internal consistency and, given that items #55 and #56 had

the least number of missing values for the Future Growth sub-scale, it was considered inappropriate to delete either item.

Similarly, the low factor loading and item statistics of item #41 cast doubt on its ability to contribute to an internally consistent description of the sub-scale. Nonetheless, as item #41 forms part of the Capacity scale, which passed its test of uni-dimensionality, and since deleting this item would not substantially raise the level of internal reliability of this sub-scale it was decided that it would be premature to exclude this item from the PI without further evidence to support this conclusion.

3.4.2 Item factor loadings for Sample B

In general, item factor loadings were satisfactory for Sample B as they varied between 0,308 and 0,773, with 93% percent (52 items) exceeding the 0,500 cut off point. In addition to item #55 discussed above, the three items that returned factor loadings below the cut off point included item #14 (0,458), item #24 (0,460) and item #39 (0,308). Interestingly, although they also appear theoretically related to the other items in the sub-scales, a clear overlap between these three items can easily be identified as they all either relate to profit and or rewards and benefits.

- Item#14 falls within the Core People Processes sub-scale and refers to *the application of rewarding unit managers and members for profits or performance, subordinate growth and creating a viable working group.* Other items in this sub-scale are quite varied and relate to: utilisation of goals and work plans, level of communication, level of decision-making, organisational interaction, conflict management, productive clashing of ideas, value placed on individual integrity and uniqueness, and learning through feedback.
- Item #24 forms part of the Employee Satisfaction sub-scale and refers to *salaries and fringe benefits.* As discussed above, other items in this sub-scale relate to satisfaction with leadership and satisfaction with work content and development.
- Item #39 refers to *the perception of profits over the past five years,* whilst other items in the Capacity sub-scale refer to the adequacy of investment in the work unit in the past and more specifically to the financial resources, quality, physical resources and assets, as well as material supply available to the unit. An additional item, item #41, refers to diversity of staff and logically stands out as different from the other items in this sub-scale. It was found to be problematic for Sample A as discussed above.

A couple of hypothesis may be put forward to explain the poor performance of these items for Sample B when compared to Sample A. On the one hand the actual wording of item #39 could have been interpreted by respondents as referring to the company's profits and not the work unit's profits, whilst other items in the Capacity sub-scale referred quite specifically to aspects of the work unit's capacity. Alternatively, item performance may have been limited due to the sampling strategy employed as Sample B consisted of only one company within the mining industry. At the time of the survey, this company had been experiencing the negative effects of an inflated gold price on operating profits. In addition, work units on mines may be seen to be somewhat removed from company profits and far more focused on operational targets, possibly because the selling price of their products (gold, platinum etc.), and thus profitability, is largely dictated by market forces.

Future studies will have to answer the above tentative explanations and provide a clearer picture of to what extent these items are influence by situational factors, and if the specific items or sub-scales should be refined. For the time being, though, there is insufficient reason to exclude any of these items, particularly because their removal would not raise the reliability coefficient substantially, and each item may be seen to operationalise an important component of the latent variables being measured by their respective sub-scale.

3.4.3 Dimensionality analysis results for Sample A

The results of the principal axis factor analysis for Sample A are summarised in Table 7. Two of the eight sub-scales failed the uni-dimensionality test, namely Employee Satisfaction and Adaptability. The problem, however, could not be solved through the deletion of individual wayward items. The Employee Satisfaction sub-scale presented clear, easily interpretable two-factor orthogonal factor structures that could be defined according to a common theme in the items loading on each factor, namely 1) a Leadership Satisfaction factor and 2) a Work Satisfaction factor (see Table 8).

The Leadership Satisfaction factor includes items that relate to the outcomes of leadership, such as respect, trust, quality of the supervision and acceptance of the leaders' influence. The Work Satisfaction factor refers to the degree to which employees appear satisfied with the task and work context, salary and fringe benefits, career development and empowerment. Factor analysis and item analysis was performed on the new factors originating from a subdivided scale. The

new factors were shown to have good internal consistency (0,900; 0,762) and all sub-divided items loaded satisfactorily on a single factor (0,511 < λ <0,729).

TABLE 7

PRINCIPLE AXIS FACTOR ANALYSES OF PI SUB-SCALE MEASURES FOR SAMPLE A

Sub-scale	KMO	% Variance explained	Min factor loading	Max factor loading
Product	0,802	42,527	0,566	0,762
Core people	0,889	39,634	0,508	0,747
Climate	0,870	52,422	0,671	0,820
	0,902	Factor 1: 35,612	0,519	0,870
Satisfaction		Factor 2: 22,633	0,499	0,775
		Single forced factor: 49,714	0,472	0,854
	0,822	Factor 1: 25,653	0,495	0,717
Adaptability		Factor 2: 23,624	0,631	0,731
		Single forced factor: 42,044	0,511	0,729
Capacity	0,855	40,543	0,422	0,781
Market standing	0,831	43,399	0,556	0,747
Future growth	0,757	40,651	0,445	0,742

TABLE 8

FACTOR LOADINGS FOR THE SATISFACTION SUB-SCALE FOR SAMPLE A (ROTATED FACTOR MATRIX)

	Factor	
	1	2
Item 22	0,373	**0,499**
Item 23	**0,519**	0,423
Item 24	0,173	**0,584**
Item 25	0,218	**0,775**
Item 26	0,432	**0,559**
Item 27	**0,852**	0,294
Item 28	**0,870**	0,259
Item 29	**0,815**	0,292
Item 30	**0,623**	0,339

The descriptive statistics calculated for the items of the Employee Satisfaction sub-scale suggest that the two factors may have emerged as an artefact of the skewness of the items. The descriptive statistics for this sub-scale for Sample A is given in Table 9. Notably, four of the five items that loaded on factor one displayed significant negative skewness whilst none of the four items that loaded on factor two displayed significant skewness. The meaningfulness of the themes shared by the items loading on the two factors on the other hand suggests that the skewness artefact hypothesis should be tempered. Despite the meaningfulness of the foregoing factor fission the Satisfaction subscale was not subdivided for the current analysis. When forcing the extraction of a single Satisfaction factor reasonably acceptable factor loadings (0,472< λ <0,854) for all items in this sub-scale were obtained.

TABLE 9

DESCRIPTIVE STATISTICS FOR THE EMPLOYEE SATISFACTION SUB-SCALE FOR SAMPLE A

	Item 22	Item 23	Item 24	Item 25	Item 26	Item 27	Item 28	Item 29	Item 30
Factor	2	1	2	2	2	1	1	1	1
N Valid	277	277	277	277	277	277	277	277	277
Mean	3,379	3,585	2,805	2,809	3,379	3,874	3,816	3,690	3,675
Std. Deviation	0,806	0,858	0,977	1,044	0,988	0,941	0,981	0,962	0,870
Variance	0,649	0,736	0,955	1,090	0,975	0,886	0,962	0,925	0,756
Skewness	0,168	-0,162	-0,023	-0,052	-0,251	-0,559	-0,505	-0,352	-0,416
Std. Error of Skewness	0,146	0,146	0,146	0,146	0,146	0,146	0,146	0,146	0,146
Kurtosis	0,219	0,067	-0,038	-0,403	-0,076	0,159	-0,136	-0,172	0,496
Std. Error of Kurtosis	0,292	0,292	0,292	0,292	0,292	0,292	0,292	0,292	0,292

In comparison, a somewhat more subtle argument needs to be mobilized in the case of the Adaptability sub-scale. A clear two-factor orthogonal factor structure was identified. Collective themes between items were not as evident. Nonetheless it could be argued that factor 1 represents a *flexibility in existing processes and structures* factor while factor 2 represents a *capacity/potency to respond efficiently to future challenges/opportunities* factor. As the rotated factor matrix for Adaptability depicted below in Table 10 illustrates, items 31-34 share the *flexibility in existing processes and structures* theme and load on factor 1 whereas items 35-37 share the *capacity to respond sufficiently to future opportunities* and load on factor 2. All items had factor loadings between 0,512 and 0,727 on a unitary factor prior to factor rotation.

TABLE 10

FACTOR LOADINGS FOR ADAPTABILITY SUB-SCALE FOR SAMPLE A (ROTATED FACTOR MATRIX)

	Factor	
	1	2
Item 31	**0,495**	0,399
Item 32	**0,717**	0,264
Item 33	**0,716**	0,225
Item 34	**0,521**	0,194
Item 35	0,400	**0,635**
Item 36	0,226	**0,731**
Item 37	0,203	**0,631**

In order to retain the simplicity of the PI model and given the two-factor model was barely distinguished from a one-factor model (second factor eigenvalue of 1,021) the Adaptability sub-scale was not subdivided into two factors. When forcing the extraction of a single Adaptability

factor reasonably satisfactory factor loadings (0,538 < λ < 0,722) for all items in this sub-scale were obtained

3.4.4 Dimensionality analysis results for Sample B

The results of the principal factor analysis for Sample B are summarised in Table 11. Three of the eight sub-scales failed the uni-dimensionality test, namely Employee Satisfaction, Capacity and Market Standing.

TABLE 11

PRINCIPLE FACTOR ANALYSES OF PI SUB-SCALE MEASURES FOR SAMPLE B

Sub-scale	KMO	% Variance explained	Min factor loading	Max factor loading
Product	0,808	45,622	0,566	0,723
Core people	0,900	38,560	0,458	0,698
Climate	0,883	53,736	0,688	0,773
	0,903	Factor 1: 37,151	0,627	0,861
		Factor 2: 22,743	0,553	0,739
Satisfaction		Single forced factor: 49,599	0,436	0,824
Adaptability	0,870	43,647	0,500	0,761
	0,810	Factor 1: 27,440	0,596	0,770
		Factor 2: 15,742	0,376	0,835
Capacity		Single forced factor: 33,540	0,294	0,743
	0,808	Factor 1: 25,650	0,439	0,709
		Factor 2: 20,338	0,471	0,690
Market standing		Single forced factor: 36,822	0,519	0,692
Future growth	0,729	39,369	0,491	0,732

Sample A and Sample B displayed almost identical patterns of factor loadings for Employee Satisfaction and Market Standing. The problem could again not be solved through the deletion of individual wayward items for these two sub-scales. As such, the Employee Satisfaction sub-scale could easily be divided into two factors, namely 1) a Leadership Satisfaction factor and 2) a Work Satisfaction factor. Despite the meaningfulness of the foregoing factor fission the Satisfaction subscale was not subdivided for the current analysis. When forcing the extraction of a single Satisfaction factor reasonably acceptable factor loadings (0,436 < λ < 0,824) for all items in this sub-scale were obtained.

The Market Standing sub-scale also presented a clear, easily interpretable two-factor orthogonal factor structure as illustrated below in Table 12 by the factor loadings. These factors could be defined according to common themes shared by the items loading on them and related to 1) Market Dominance which included market share, competitiveness in markets, and diversity of

markets or products, and 2) Reputation, including competitiveness, customer satisfaction and reputation for adding value. Factor and item analyses were performed on the new factors originating from the subdivided scales. The new factors were shown to have reasonable internal consistency (0,732; 0,703) although if item #49 was deleted it would raise the Cronbach alpha of factor 1 slightly (0,010). All sub-divided items loaded satisfactorily on a single factor (factor 1: $0{,}439 < \lambda < 0{,}709$; factor 2: $0{,}471 < \lambda < 0{,}690$).

TABLE 12

FACTOR LOADINGS FOR MARKET STANDING SUB-SCALE FOR SAMPLE B (ROTATED FACTOR MATRIX)

	Factor	
	1	**2**
Item 45	**0,610**	0,334
Item 46	**0,709**	0,244
Item 47	**0,678**	0,098
Item 48	0,436	**0,471**
Item 49	**0,439**	0,269
Item 50	0,187	**0,690**
Item 51	0,205	**0,688**

Despite the meaningfulness of the foregoing factor fission the Market Standing subscale was not subdivided for the current analysis. When forcing the extraction of a single Market Standing factor reasonably acceptable factor loadings ($0{,}519 < \lambda < 0{,}692$) for all items in this sub-scale were obtained.

The Capacity sub-scale presented a somewhat more complex problem to resolve. The factor loadings for the Capacity sub-scale for Sample B are summarised in Table 13. The two-factor model was easily distinguished from a one-factor model (second factor eigenvalue of 1,120). The two factors seem to distinguish between the wealth of human resource resources and the wealth of financial and physical resources. Factor 1 seems to represent a *Wealth of financial and physical resources factor* while factor 2 seems to represents a *Wealth of human resources factor.* Although loading primarily on factor 2, item #39 did nonetheless not load strongly on this factor. This item assesses the extent to which profits showed growth over the past five year. The fact that the item loaded moderately on factor 2 suggests that profit growth is perceptually associated with (or attributed to) the human resource capacity of the unit. Despite the meaningfulness of the foregoing factor fission the Capacity subscale was not subdivided for the current analysis. When forcing the extraction of a single Capacity factor reasonably acceptable factor loadings ($0{,}453 < \lambda < 0{,}743$) for all items in this sub-scale were obtained but for item #39 ($\lambda = 0{,}294$).

TABLE 13

FACTOR LOADINGS FOR CAPACITY SUB-SCALE FOR SAMPLE B

(ROTATED FACTOR MATRIX)

	Factor	
	1	**2**
Item 38	**0,599**	0,190
Item 39	0,127	**0,376**
Item 40	0,143	**0,835**
Item 41	0,367	**0,402**
Item 42	**0,770**	0,176
Item 43	**0,670**	0,203
Item 44	**0,596**	0,180

3.4.5 Overall skewness

Descriptive item statistics for Sample A and Sample B are provided as Appendix 1 and Appendix 2. The majority of items followed a negatively skewed and leptokurtic distribution. More items in Sample A followed a significantly ($p<0,05$) negatively skewed distribution (18 items, 32%) when compared to Sample B for which 9 items (16%) displayed significant skewness, with 4 (7%) positively skewed and 5 (9%) negatively skewed. These differences between the distributions for Sample A and Sample B suggest that the work units in Sample A were generally evaluated more positively with relatively few units being poorly evaluated when compared to Sample B where a somewhat more balanced evaluation of units were obtained. Interestingly this may be explained by the fact that the majority of Sample A constituted part-time MBA students who tend to be high performers who are rewarded for good performance by their respective companies by being sponsored to do their MBAs. Sample B included managers across a single company who attended a management development programme. Almost all managers at a certain level within this company attend this programme, and therefore the sample is likely to show a balanced distribution between good and poorly performing work units.

3.4.6 Discussion on the item- and dimensionality analyses

Comparing the dimensionality analyses results for Sample A and Sample B shows support for the hypotheses of possible factor fission across the Employee Satisfaction, Adaptability, Capacity and Market Standing sub-scales that was identified in previous research (Henning *et al.,* 2003; Theron *et al.,* 2004). Table 14 provides a dimensionality comparison between the sub-scales of Sample A and Sample B that highlights this pattern. In this study, factor fission was found to result in a

conceptually meaningful division of Employee Satisfaction and Adaptability for Sample A, and Employee Satisfaction, Capacity and Market Standing for Sample B. Table 14, however, also suggests that the uni-dimensionality assumption could have been rejected in Sample A with regards to Capacity and Market Standing if the eigenvalue greater than unity extraction rule had been slightly relaxed. Moreover the same factor pattern to that found in Sample B would most probably appear in Sample A with regards to these two sub-scales if the extraction of two factors would have been forced. In other words, examining the PI via different data sets essentially reveals different aspects of one reality, namely that a theoretically meaningful refinement of the unit performance model may be possible by splitting Market Standing, Satisfaction, Capacity and Adaptability. For the purposes of this study, however, the original unit performance dimensions will not be extended as doing so would further complicate an already complex model. It moreover, would defeat the ultimate purpose of the measurement and structural invariance studies namely to evaluate the generalizability of the basic PI measurement and structural model (Henning *et al.*, 2003) across random samples from the PI target population. If the hypothesised measurement model satisfactorily fits the data, it would support subsequent research which should investigate further refinements suggested in the foregoing results.

TABLE 14

DIMENSIONALITY COMPARISON BETWEEN SAMPLE A AND SAMPLE B

		Initial Eigenvalues per sub-scale							
Sample	Factor	Production and Efficiency	Core People Processes	Climate	Employee Satisfaction	Adaptability	Capacity	Market Standing	Future Growth
A	1	2,681	4,154	4,137	4,928	3,415	3,379	3,589	2,572
	2	0,702	0,764	0,895	1,089	1,021	0,916	0,981	0,825
B	1	2,813	4,069	4,222	4,918	3,597	2,955	3,194	2,547
	2	0,756	0,899	0,802	1,272	0,901	1,120	1,056	0,912

In summary, although no conclusive evidence in this regard can be derived from the current samples, the foregoing analyses indicate that the PI items, for the most part, systematically reflect their designated latent unit performance dimensions with reasonable success and do not reflect artefact factors or an extensive amount of non-relevant information. Results on the fit of the first-order measurement model for both Sample A and Sample B reported below tend to increase the confidence in this position. To more convincingly substantiate the position that the PI successfully measures the performance construct as it had been constitutively defined, the nomological network in which the performance construct is imbedded would have to be evaluated via structural equation modelling.

3.5 VARIABLE TYPE AND ITEM PARCELLING

Structural equation modelling (SEM) was used to perform a confirmatory factor analysis on the reduced data sets for both Sample A and Sample B, obtained after imputation of missing values. For this purpose, two indicator variables (item parcels) were created from each sub-scale. The process of creating the indicator variables is discussed below. The interest in applying parcels within SEM is largely based on its proposed advantages compared to single items. These advantages are either related to the difference in psychometric characteristics between items and parcels or related to factor-solution and model-fit advantages accruing to models based on parcels (Little *et al.*, 2002).

3.5.1 Difference in psychometric characteristics between items and parcels

Item parcelling has the potential to serve as a data analysis panacea for a variety of data problems, primarily non-normality, small sample sizes, small sample size to variable ratio, and unstable parameter estimates (Bandalos *et al.*, 2001). As the most frequently used estimators in SEM require normally distributed continuous variables, item parcels have been preferred over single items as indicators of latent constructs because they better approximate normally distributed continuous variables if used as indicators of latent constructs (Bentler & Chou, 1987). By creating parcels, researchers can construct new variables that are closer to being continuous (better approximations to normally distributed continuous variables), which allows for a distribution closer to normal, and may therefore reduce distortion of estimates (Bandalos, 2002). Therefore parcels are more likely to meet the assumptions of Maximum Likelihood estimation than are individual ordered-categorical items. In other words parcelling can be viewed as a heuristic approach to converting ordered-categorical data to continuous data with an eye toward minimising the attenuation caused by using ordered-categorical variables (Nasser & Takahashi, 2003).

3.5.2 Factor-solution and model-fit advantages and disadvantages

Additional advantages of item parcels include that the composite score of an item parcel is normally more reliable than single item scores. Item parcels also yield variance-covariance matrices that are amenable to linear factor analysis (Hagvet & Nasser, 2004). Nasser and Takahashi (2003) also found that lower skewness and kurtosis and higher validity occur for

parcels than for individual items; specifically item parcels with more than two items were found to exhibit less skewness and kurtosis and higher reliability and validity (Marsh, Hau, Balla, & Grayson, 1998). When compared with individual items, model-fit indices as measured by the root means squared error or approximation (RMSEA), comparative fit index (CFI), and the chi-square test also improve systematically as the number of items per parcel increased, provided items had a uni-dimensional structure (Bandalos, 2002).

3.5.3 Potential disadvantages of item parcelling

Although it has many advantages, aggregating information from items into parcels is not without its problems. As Holt (2003) and Little *et al.* (2002) emphasise, item parcels work best when constructed on uni-dimensional structures. Item parcels drawn from items assessing a multi-dimensional construct are themselves likely to be multidimensional in composition, leading to difficulties in interpretation. An additional concern is that item parcelling may improve model fit for all models, even if they are misspecified (Bandalos *et al.,* 2001). This is because parcel-based models tend to cancel out random and systematic error by aggregating across these errors, thereby improving model fit. As a result item parcelling may reduce the probability that misspecified models may be identified, and thus possibly increase the chance of Type II errors (failing to reject a model that should have been rejected) (Little *et al.,* 2002).

Nonetheless, there are very few solutions besides item parcelling or item-based analysis that may be used in this current research. One such solution would be to create a subset of observed variables for model fitting and testing. However, this is likely to result in some variables being discarded from a CFA. These eliminated variables may contain valuable information on estimators and tests concerning the structural model (Nasser & Takahashi, 2003), which eliminated it as an option in this study. Rather, item parcelling was considered the better option.

3.5.4 Appropriateness of using item parcelling for this research

Although individual PI items should be considered ordinal variables as a five-point Likert scale is used to capture responses, SEM on the PI in which each individual item serves as a manifest or indicator variable of the various latent unit performance facets would have resulted in a cumbersome and extensive exercise simply due to the number of items involved. The ordinal nature of the data would have required the asymptotic covariance or asymptotic variance matrices

to be calculated which demands extensive memory and processing time when the number of variables are large (Jöreskog & Sörbom, 1996b). In a cross-validation study such as this research, the comparison between two samples using individual items rather than parcels would exacerbate this problem. Therefore it was decided to use items parcels for this research, but to remain alert to the potential consequences of doing so when considering model fit.

3.5.5 Generating item parcels

Although item parcelling is frequently used, there is no consensus amongst researchers on how items should be aggregated into parcels. Nonetheless, research indicates that how well parcels work depends to a large extent on the specific allocations of items into parcels (Holt, 2003; Little *et al.,* 2002). There appear to be two main considerations when allocating items into parcels. According to Hagvet and Nasser (2004) parcels are acceptable indicators of the latent construct if 1) they meet the parametric assumptions for uni-dimensionality, and 2) if items and parcels have content validity as measures of the latent construct. The following recommendations based on Holt (2003), Hagvet and Nasser (2004), Hagvet and Zuo, (2000), Little *et al. (*2002), and Hall, Snell and Foust (1999), were considered when constructing item parcels for this research:

Recommendation 1: Check for uni- or multi-dimensionality of factors

Many authors recommend that item parcelling should only be used if the items to be parcelled are from a uni-dimensional sub-scale. However, there is some argument for using item parcels regardless of their dimensionality. As mentioned above, the number of items in the PI would make an analysis cumbersome if they were used as individual indicator variables. In addition, there is evidence that the maximum likelihood (ML) method employed in this research cannot provide a reliable inference when the number of variables in an analysis becomes excessively large. An excessive number of variables especially in relation to sample size are likely to result in misleading findings and invalid conclusions regarding the factor structure (Bernstein, Teng, Grannemann & Garbin, 1987; Kishton & Widaman, 1994, cited in Nasser & Takahashi, 2003, p. 76).

Furthermore, Hall *et al. (*1999) propose using sub-scale dimensionality in item parcelling for a different reason. If multi-dimensional sub-scales are identified, isolated parcelling strategies could be used to capture similar facets of the structure into the same item parcel and different facets would be separated into different parcels. In a comprehensive study on different item

parcelling strategies, Hall *et al. (*1999) found that generating item parcels in this way forces any unmodelled influence into the uniqueness terms, and thereby isolates the factor loadings and structural estimates as much as possible from the contamination of the secondary factor influence. In contrast, when parcels are created without allocating items according to multi-dimensional nature of the sub-scale, parcelling can obscure the true factor structure and result in biased parameter estimates and inflated indices of fit (Hall *et al.,* 1999).

Recommendation 2: Consider the normality and difficulty of the items

Secondly, the normality/difficulty of the original items to be parcelled was important to consider as it was recommended that items with very non-normal distributions should be combined with other items in such a way as to maximise the normality of the resulting parcels. As the PI has ordered, categorical items, this can be accomplished by combining items with opposite skew or distributional shape (Holt, 2003).

Recommendation 3: Check content validity of parcels

Thirdly, conducting a conceptual analyses of item content enables the researcher to check that items and parcels have content validity as measures of the latent construct (Hagvet & Zuo, 2000), and that they represent the facets underlying the dimension as far as possible (Little *et al.,* 2002).

Recommendation 4: Create the least number of parcels with the most items.

Finally, it is recommended to create a limited number of parcels from the items available for each latent variable rather than numerous parcels containing only a limited number of items. As the number of items in each of the eight PI sub-scales ranges from five to nine, it does not seem appropriate to create more than two item parcels for each sub-scale.

3.5.6 Generating item parcels based on recommendations

Based on the above recommendations, the following process was followed when forming item parcels. Item parcelling was based on the item analysis and dimensionality analysis results obtained for Sample B, but checked against the results obtained for Sample A to ensure no contradictions were evident. In the first step, all sub-scales were first subdivided according to alternate allocations of items according to highest to lowest factor loadings. As discussed above, Employee Satisfaction and Market Standing could easily be divided into two item parcels representing conceptually meaningful division of the original sub-scale in question. However,

item #39 of the Capacity sub-scale proved to be problematic and could not be easily allocated to either of the clearly identifiable factors and therefore items were allocated as though Capacity was uni-dimensional. The second step involved comparing the skewness of items in each parcel from a uni-dimensional sub-scale in order to ensure that items with very non-normal distributions were balanced out between the parcels for each sub-scale. This did not pose a problem for Sample B or Sample A. It was not considered problematic that an item parcel from each of the Market Standing and in Employee Satisfaction sub-scales contained many significantly negatively skewed items for Sample A as both sub-scales were multi-dimensional. The last step involved considering if any items were conceptually ill-fitted to the item parcel they were allocated to. There was no need to change the item allocations.

Following the above process, two indicator variables were created from each sub-scale and are presented in Table 15. These item parcels reduced the manifest variables in the measurement model from fifty-six to sixteen. These composite indicator variables were treated as continuous variables, thus allowing for the analysis of the covariance matrix (Jöreskog & Sörbom, 1996b). In the context of multi-group analyses, this sampling strategy was quite tricky to implement. For example, it is hard to justify why the results of the dimensionality analysis for Sample B were used rather then those of Sample A. Nonetheless, by using this complex item parcelling strategy it was hoped that more reliable indicator variables would be created, and that this would enhance the credibility of this study's results.

TABLE 15

ITEM – PARCEL ALLOCATIONS FOR SAMPLE A AND SAMPLE B

Sub-scale	Items allocated to Parcel 1	Items allocated to Parcel 2
Production and Efficiency	2; 3; 4	1; 5
Core people processes	8; 10; 12; 13; 14	6; 7; 9; 11;
Work unit climate	16; 18; 19; 20	15; 17; 21;
Employee Satisfaction	23; 27; 28; 29; 30	22; 24; 25; 26
Adaptability	31; 32; 34; 37	33; 35; 36
Capacity	38; 39; 41; 42	40; 43; 44
Market Standing	45; 46; 47; 49	48; 50; 51
Projected Future Growth	52; 54; 55	53; 56

3.6 UNIVARIATE AND MULTIVARIATE NORMALITY

Maximum likelihood is considered as the preferred method of estimation when fitting measurement models to continuous data. However, maximum likelihood assumes multivariate normality, as does the alternative estimation methods for structural equation modelling with continuous data, namely true generalised least squares (GLS) and full information maximum likelihood (FIML) (Mels, 2003). Although the assumption of multivariate normality may in some cases of single populations be justifiable, it is more complicated in a multigroup context. As Lubke and Muthen (2004, p. 515) state "results from robustness studies in a single homogeneous population concerning the analysis of Likert-type data while violating the normality assumption, do not necessarily carry over to the multiple group situation, and group comparisons may have problems in addition to those encountered in single populations". In an empirical study, Lubke and Muthen (2004) found the source of unacceptable fit remains obscure in multigroup CFA of ordered categorical data whilst incorrectly assuming multivariate normality. They concluded that a researcher would not know whether unfavourable measures of goodness-of-fit are really due to a violation of ML assumptions, due to threshold differences across items that result in structural differences, or due to the fact that the data are categorical and measures of GIF based on the assumption of normally distributed data do not function properly. Inappropriate analysis of continuous non-normal variables in SEM can moreover result in incorrect standard errors and chi-square estimates (Du Toit & Du Toit, 2001).

The univariate and multivariate normality of the composite indicator variables of the PI were evaluated by means of PRELIS (Jöreskog & Sörbom, 1996b). The null hypothesis of univariate normality had to be rejected for five of the sixteen composite indicator variables for Sample A, and for six composite indicator variables for sample B. This is consistent with the skewness findings reported earlier. The results of the test for multivariate normality are given in Table 16. The assumption of multivariate normality is evidently not defensible in both samples A and B.

TABLE 16

TEST OF MULTIVARIATE NORMALITY FOR CONTINUOUS VARIABLES

		Skewness			Kurtosis			Skewness and Kurtosis	
Sample	N	Value	Z-Score	P-Value	Value	Z-Score	P-Value	Chi-Square	P-Value
A	277	27,150	9,338	0,000	315,898	7,520	0,000	143,736	0,000
B	375	20,804	10,197	0,000	318,739	9,208	0,000	188,779	0,000

Two possible solutions to the lack of normality were explored. The first solution involved normalising the composite indicator variables using PRELIS. Although the skewness and kurtosis of the indicator variable distributions significantly improved when normalised, the null hypothesis of multivariate normality still had to be rejected for both samples as shown in Table 17. This impasse occurs because indicator variables are normalised independently of each other.

TABLE 17

TEST OF MULTIVARIATE NORMALITY FOR NORMALISED CONTINUOUS VARIABLES

		Skewness			Kurtosis			Skewness and Kurtosis	
Sample	N	Value	Z-Score	P-Value	Value	Z-Score	P-Value	Chi-Square	P-Value
A	277	24,903	7,353	0,000	308,704	6,121	0,000	91,537	0,000
B	375	21,070	10,497	0,000	319,606	9,384	0,000	198,253	0,000

As the normalised indicator parcels did not pass the test of multivariate normality, the second solution involved exploring alternative methods of estimation that are better suited to data that does not follow a multivariate normal distribution. Methods that are recommended to fit structural equation models to non-normal data include, robust maximum likelihood (RML), weighted least squares (WLS) and diagonally weighted least squares (DWLS) (Mels, 2003). Such methods provide an advantage over the use of normal scores in that solutions do not have to be interpreted in terms of transformed values (Du Toit & Du Toit, 2001). For this study, robust maximum likelihood was selected as the method of estimation based on Mels (2003) recommendation that it be used if the assumption of a multivariate normal distribution does not hold.

CHAPTER 4
EVALUATION OF THE MEASUREMENT MODEL

4.1 THE PI MEASUREMENT MODEL

The PI was developed to measure the multifaceted construct of work unit performance. The facets of the work unit performance construct were purposefully conceptualized in terms of eight unit performance dimensions to which specific constitutive meanings have been attached as shown in Table 1 (see page 7). These unit performance latent dimensions were subsequently purposefully operationalized in terms of specific effect indicators (Babbie & Mouton, 2001). Specific items were written to function as relatively uncontaminated behavioural expressions of each latent work unit performance dimension. Measurement model fit for independent samples indicates the degree to which the researchers have succeeded in their measurement intentions. As such, poor measurement model fit would question the extent to which the PI's operational design is able to provide a comprehensive and uncontaminated empirical grasp on the work unit performance construct as defined. It makes sense, therefore, to first establish whether the model fits on both samples independently before proceeding to investigate whether the parameter estimates can be considered equal across samples though tests of measurement invariance.

Structural equation modelling (SEM) was used to perform a confirmatory first-order factor analysis on the parcelled data sets for each sample (Kaplan, 2000). The conceptualisation of the unit performance construct, in conjunction with the architecture to the PI, implies a specific factor structure or measurement model. Given that two manifest variables were created from each sub-scale, the measurement model underlying the PI can be shown in matrix format as equation 1.

$$X = \Lambda^{x}\xi + \delta \qquad 1$$

Where:

- X is a 16x1 column vector of observable indicator scores;
- Λ^{x} is a 16x8 matrix of factor loadings;
- ξ is a 8 x 1 column vector of first-order latent unit performance facets; and

- δ is a 16x1 column vector of unique/measurement error components comprising the combined effect on X of systematic non-relevant influences and random measurement error (Jöreskog & Sörbom, 1996b).

The measurement model implies two additional matrices. A symmetric 16x16 covariance/correlation matrix Φ contains the correlations between the latent unit performance dimensions. A diagonal 16x16 matrix θ_δ depicts the variance in the error terms associated with the indicator variables. The diagonal nature of the θ_δ matrix implies that the error terms δ are assumed to be uncorrelated across the indictor variables. If the measurement model would make provision for correlated error terms by freeing the off-diagonal elements of θ_δ, it would imply the existence of additional common factors, not reflected in the model, but which also underlie the response to the indicator variables (or possibly causal effects existing between the systematic error contained in the measurement error terms). No substantive justification could in this case be found to free the off-diagonal elements of θ_δ.

4.2 MODEL IDENTIFICATION

Model identification needs to be examined prior to confronting the model with data. Broadly speaking this involves determining whether one has sufficient information to obtain a unique solution for the parameters to be estimated in the model (Diamantopoulos & Siguaw, 2000). In other words model identification considers if the nature of the model and the data would permit the determination of unique estimates for the freed parameters in the model. This would be possible if for each free parameter there would exist at least one algebraic function that expresses that parameter as a function of sample variances/covariance terms (MacCallum, 1995). There is, however, no such set of necessary and sufficient conditions that if satisfied would ensure that the model is identified. At the very least, though, the following two important requirements have to be met (Diamantopoulos & Siguaw, 2000; MacCallum, 1995). Firstly, a definite scale has to be established for each latent variable. Secondly, the number of model parameters to be estimated may not exceed the number of unique variance/covariance terms in the sample observed covariance matrix (Diamantopoulos & Siguaw, 2000; MacCallum, 1995). The measurement model depicted as equation 1 satisfies both these requirements. The first requirement is met by treating each latent variable as a (0; 1) standardized variable (MacCallum, 1995). The number of model parameters that are set free to be estimated (t=60) are also less than the number of non-redundant elements in the observed sample covariance matrix ([(p+q)(p+q+1)]/2=136) whereby

p=the number of y-variables and q=the number of x-variables (Diamantopoulos & Siguaw, 2000). This results in the rather moderate degrees of freedom of 76.

4.3 INDEPENDENT ASSESSMENT OF OVERALL GOODNESS-OF-FIT OF THE FIRST-ORDER MEASUREMENT MODEL FOR SAMPLE A AND SAMPLE B

Prior to cross-validating the measurement model, it is necessary to fit the model on both samples independently. Poor model fit for either one of the samples or both would cast doubt on the value of cross-validating the measurement model across these two specific samples. LISREL 8.53 (Du Toit & Du Toit, 2001; Jöreskog & Sörbom, 1996b) was used to determine the fit of the PI model shown as equation 1. For the purposes of confirmatory factor analysis, the measurement model was treated as an exogenous model, simply for programming reasons. The data was first read into PRELIS (Jöreskog & Sörbom, 1996b) to compute covariance and asymptotic covariance matrices to serve as input for the LISREL analysis. The model fit was evaluated through an analysis of a covariance matrix due to the assumed continuous nature of the item parcels. Robust maximum likelihood was used to estimate the parameters set free in the model due to the failure of the data to satisfy the multivariate normality assumption.

No single measure of fit can provide a conclusive verdict on model fit (Bollen & Long, 1993; Schumacker & Lomax, 1996). Evaluation of model fit should rather be determined through an integrative process that considers a variety of sources and is based on several criteria that assess model fit from different perspectives (Diamantopoulos & Siguaw, 2000; p. 82). The full spectrum of indices provided by LISREL to assess the absolute and comparative fit of the proposed measurement model were therefore used to reach an informed decision concerning the model's overall fit (Diamantopoulos & Siguaw, 2000). The results of the model fit analyses are presented and discussed first for Sample A and then for Sample B.

4.4 RESULTS FOR SAMPLE A

4.4.1 Overall fit assessment for Sample A

The full spectrum of indices provided by LISREL to assess the absolute and comparative fit of the proposed measurement model with the data from Sample A is presented in Table 18. An admissible final solution of parameter estimates for the PI measurement model was obtained after 9 iterations.

TABLE 18

GOODNESS-OF-FIT INDICATORS FOR SAMPLE A

Degrees of Freedom = 76
Minimum Fit Function Chi-Square = 141,42 (P = 0,00)
Normal Theory Weighted Least Squares Chi-Square = 140,04 (P = 0,00)
Satorra-Bentler Scaled Chi-Square = 128,83 (P = 0,00015)
Chi-Square Corrected for Non-Normality = 156,95 (P = 0,00)
Estimated Non-centrality Parameter (NCP) = 52,83
90 Percent Confidence Interval for NCP = (25,35 ; 88,18)

Minimum Fit Function Value = 0,51
Population Discrepancy Function Value (F0) = 0,19
90 Percent Confidence Interval for F0 = (0,092 ; 0,32)
Root Mean Square Error of Approximation (RMSEA) = 0,050
90 Percent Confidence Interval for RMSEA = (0,035 ; 0,065)
P-Value for Test of Close Fit (RMSEA < 0,05) = 0,47

Expected Cross-Validation Index (ECVI) = 0,90
90 Percent Confidence Interval for ECVI = (0,80 ; 1,03)
ECVI for Saturated Model = 0,99
ECVI for Independence Model = 31,72

Chi-Square for Independence Model with 120 Degrees of Freedom = 8723,93
Independence AIC = 8755,93
Model AIC = 248,83
Saturated AIC = 272,00
Independence CAIC = 8829,92
Model CAIC = 526,27
Saturated CAIC = 900,87

Normed Fit Index (NFI) = 0,98
Non-Normed Fit Index (NNFI) = 0,99
Parsimony Normed Fit Index (PNFI) = 0,62
Comparative Fit Index (CFI) = 0,99
Incremental Fit Index (IFI) = 0,99
Relative Fit Index (RFI) = 0,97

Critical N (CN) = 210,97

Root Mean Square Residual (RMR) = 0,015
Standardized RMR = 0,029
Goodness-of-fit Index (GFI) = 0,94
Adjusted Goodness-of-fit Index (AGFI) = 0,89
Parsimony Goodness-of-fit Index (PGFI) = 0,53

The chi-square test statistic indicates whether the observed and estimated covariance matrices differ relative to sample size (Pousette & Hanse, 2002). The Satorra Bentler chi square results from the use of robust maximum likelihood which has been chosen because it is better suited to multivariate non-normal data (Diamantopoulos & Siguaw, 2000). The Satorra-Bentler χ^2 test statistic (128,83) is significant ($p<0,01$) thus resulting in a rejection of the null hypothesis of exact model fit (H_0: $\Sigma=\Sigma(\theta)$). This means the first-order measurement model is not able to reproduce the observed covariance matrix to a degree of accuracy that could be explained in terms of sampling error only. This result, however, is not surprising because the χ^2 value is used to

determine if significant variance is left unexplained, and significant unexplained variance is typical in social science research (Peterson *et al.,* 1995). As such, the null hypothesis that the model fits perfectly to the population may be considered to be rather unrealistic (Browne & Cudeck, 1993). Further difficulties of using the chi-square statistic include that it is sensitive to sample size (Diamantopoulos & Siguaw, 2000), and there is also no consensus about what χ^2 value represents a good fit, even though the χ^2 measure is the only statistically-based measure of goodness-of-fit available in SEM (Bollen, 1989, cited in Poulette & Hanse, 2002, p. 231; Hu & Bentler, 1995).

It is, therefore, recommended that the chi-square statistic should be treated as a descriptive badness-of-fit measure (Pousette & Hanse, 2002). This can be done by using the normed χ^2 measure to identify inappropriate models. Normed χ^2 values less then 1,0 indicate an 'overfitted' model (Schumacker & Lomax, 1996) whilst ratio values more than 2,0 (or the more liberal limit of 5,0) indicate that the model does not fit the observed data and needs improvement (Pousette & Hanse, 2002). For sample A, the normed χ^2 expressed as the Satorra-Bentler χ^2 estimate in terms of the degrees of freedom (χ^2/df = 1,69) suggests that the measurement model demonstrates acceptable fit to the data. Kelloway (1998), though, advises against a strong reliance on the normed χ^2 as the guidelines indicative of good fit have very little empirical justification.

In most circumstances the hypothesised model is only an approximation to reality which means the χ^2 test statistic will follow a non-central χ^2 distribution with non-centrality parameter, λ. The estimated λ assesses the degree of model fit by estimating the discrepancy between the observed (Σ_0) and estimated population covariance ($\widetilde{\Sigma}_0$). The larger the estimated λ, the farther apart is the true alternative from the null hypothesis (Diamantopoulos & Siguaw, 2000). For Sample A, the estimated λ value (52,83) is not very high. The 90 percent confidence interval for NCP (25,35 ; 88,18) also ranges across acceptable values[1]. This suggests that the estimated discrepancy between the observed (Σ_0) and estimated population covariance ($\widetilde{\Sigma}_0$) matrices is not very high, which indicates good model fit (Diamantopoulos & Siguaw, 2000).

Root mean square error of approximation (RMSEA) focuses on the discrepancy between Σ and $\Sigma(\theta)$ per degree of freedom. It is generally regarded as one of the most informative fit indices,

[1] The researcher, however, has to confess that she would be rather hard-pressed to indicate exactly when the non-centrality parameter values would become unacceptable.

especially as it takes the model complexity into account (Diamantopoulos & Siguaw, 2000). The RMSEA value indicates how well the model with unknown but optimally chosen parameter values would fit the population covariance matrix if it were available. The RMSEA value (0,05) for Sample A meets the criteria of close or good model fit which is indicated by a value ≤0,05 according to the guidelines provided by Browne and Cudeck (1993). Likewise, the 90 percent confidence interval for RMSEA shown in Table 18 (0,035 ; 0,065) indicates that the fit of the structural model could be regarded as good. In addition, a test of close fit performed by LISREL shows a high probability (0,47) of obtaining a RMSEA value of 0,05 in the sample given the assumption that the model fits closely in the population. In symbol form therefore:

$$P[RMSEA=0,05 \mid H_0: RMSEA \leq 0,05 \text{ is true}]=0,47$$

The null hypothesis that RMSEA is ≤ 0,05 can therefore not be rejected. These results, therefore, all suggest that close model fit has been achieved.

The expected cross-validation index (ECVI) expresses the difference between the reproduced sample covariance matrix ($\hat{\Sigma}$) derived from fitting the model on the sample at hand and the expected covariance matrix that would be obtained in an independent sample of the same size from the same population (Byrne, 1998; Diamantopoulos & Siguaw, 2000). It focuses on overall error and is, therefore, a useful indicator of a model's overall fit (Diamantopoulos & Siguaw, 2000). Since the model ECVI (0,90) is far smaller than the value obtained for the independence model (31,72) and also less than the ECVI value associated with the saturated model (0,99), the fitted model appears to have the greatest potential of being replicated in a cross-validation sample which indicates good model fit (Diamantopoulos & Siguaw, 2000).

Indices of parsimonious fit refer to the benefit that amounts in terms of the extent to which fit is improved whilst taking into account the cost incurred in terms of degrees of freedom lost in order to attain this improved fit (Jöreskog & Sörbom, 1993). In other words, the assessment of parsimonious fit acknowledges that model fit can always be improved by adding more paths to the model and estimating more parameters until perfect fit is achieved in the form of a saturated or just-identified model with no degrees of freedom (Kelloway, 1998). The key is, therefore, to find the most parsimonious model that achieves satisfactory fit with as few model parameters as possible (Jöreskog & Sörbom, 1993). The parsimonious normed fit index (PNFI = 0,62) and the parsimonious goodness-of-fit index (PGFI = 0,53) approach model fit from this perspective. PNFI and PGFI range from 0 to 1, with higher values indicating a more parsimonious fit, however, neither index is likely to reach the 0,90 cutoff used for other fit indices and there is also

no standard for how high either index should be to indicate parsimonious fit (Kelloway, 1998). They are more inclined to be meaningfully used when comparing two competing theoretical models and, therefore, not very useful indicators in this analysis (Kelloway, 1998). Acceptable values for the PGFI according to Diamantopoulos and Siguaw (2000) generally tend to be somewhat more conservative even when other indices indicate acceptable fit.

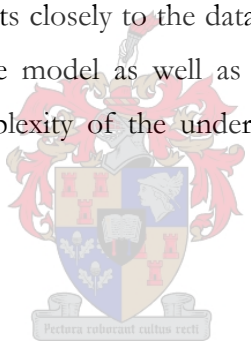
The values for the Aiken information criterion (AIC= 248,83) suggest that the fitted measurement model provides a more parsimonious fit than the independent/null model (8755,93) as well as the saturated model (272,00) (Kelloway, 1998). Similarly, the values for the consistent Aiken information criterion (526,27) imply that the fitted measurement model provides a more parsimonious fit than both the independent/null model (8829,92) and the saturated model (900,87). This, in conjunction with the ECVI results, indicates that the measurement model does not lack influential paths and, as such, may be considered the most parsimonious fit.

Relative fit indices compare the ability of the model to reproduce the observed covariance matrix with that of a baseline model, usually the independence model that postulates no paths between the variables in the model (Diamantopoulos & Siguaw, 2000). The closer the values are to 1 from 0, the better the fit with 0,90 generally considered indicative of a well fitting model (Diamantopoulos & Siguaw, 2000; Kelloway, 1998). The indices of relative fit presented in Table 18 include the normed fit index (NFI = 0,98), the non-normed fit index (NNFI = 0,99), the comparative fit index (CFI = 0,99), the incremental fit index (IFI = 0,99), and the relative fit index (RFI = 0,97). These aforementioned indices all exceed the critical value of 0,90 and therefore indicate good comparative fit relative to the independence model.

The critical sample size statistic (CN) refers to the size that the sample would have to reach in order to accept the χ^2 statistic as significant at the 0,05 significance level (Diamantopoulos & Siguaw, 2000). The estimated CN value (210,97) falls above the recommended threshold value of 200 which is regarded as indicative of the model providing an adequate representation of the data (Diamantopoulos & Siguaw, 2000) although this proposed threshold should be used with caution (Hu & Bentler, 1995). The standardized RMR may be considered a summary measure of standardized residuals which represents the average difference between the elements of the sample covariance matrix and the fitted covariance matrix. If the model fit is good the fitted residuals ($S - \Sigma\hat{}$) should be small in comparison to the magnitude of the elements in S

(Diamantopoulos & Siguaw, 2000). The RMR (0,015) and standardized RMR (0,029) also indicate good fit as values less than 0,05 on the latter index suggest the model fits the data well (Kelloway, 1998).

The goodness-of-fit index (GFI) and the adjusted goodness-of-fit index (AGFI) reflect how closely the model comes to perfectly reproducing the sample covariance matrix (Diamantopoulos & Siguaw, 2000). The AGFI (0,89) adjusts the GFI (0,94) for the degrees of freedom in the model (Diamantopoulos & Siguaw, 2000; Jöreskog & Sörbom, 1993) and should be between zero and unity with values exceeding 0,9 indicating the model fits well with the data (Jöreskog & Sörbom, 1993; Kelloway, 1998). Evaluating the fit of the model in terms of these two indices thus supports model fit, although AGFI index of 0,89 is slightly lower than the cutoff value. Kelloway (1998) states that GFI and AGFI should be used with some circumspection as guidelines for the interpretation are grounded in experience and therefore somewhat arbitrary.
In sum, when the abovementioned model fit statistics are considered in unison, they seem to unanimously suggest that the model fits closely to the data of Sample A. In addition, the model clearly outperforms the independence model as well as the saturated model and therefore it seems to fully capture the true complexity of the underlying PI model without the need for additional paths.



4.4.2 Examination of residuals

Residuals represent the differences between corresponding cells in the observed and fitted covariance matrices (Diamantopoulos & Siguaw, 2000; Jöreskog & Sörbom, 1993). As such, residuals, and especially standardized residuals, provide valuable diagnostic information on sources of lack of fit in models (Jöreskog & Sörbom, 1993; Kelloway, 1998). Standardized residuals can be interpreted as standard normal deviates (i.e. z-scores). Large positive and negative standardized residuals with absolute values greater than 2,58 would be indicative of relationships (or the lack thereof) between indicator variables that the model fails to explain (Diamantopoulos & Siguaw, 2000). Large positive residuals would indicate that the model underestimates the covariance between two observed variables. Adding paths to the model that could account for the covariance should, therefore, rectify the problem. Conversely, large negative residuals would indicate that the model overestimates the covariance between specific observed variables. Rectifying this situation would, therefore, lie in removing some or all of the

paths that are associated with the indicator variables in question (Diamantopoulos & Siguaw, 2000; Kelloway, 1998).

The stem-and-leaf plot depicted in Figure 2 confirms the positive conclusion on model fit that was suggested by the fit statistics earlier as the distribution of standardized residuals appears to be distributed approximately symmetrical around a median standardized residual of zero. The smallest (-2,02) and largest (2,49) standardized residual also fall well within the 0,01 significance limits. Overall, the absence of large positive and negative residuals suggest that the observed covariance terms in the observed sample covariance matrix are estimated reasonably well by the derived model parameter estimates.

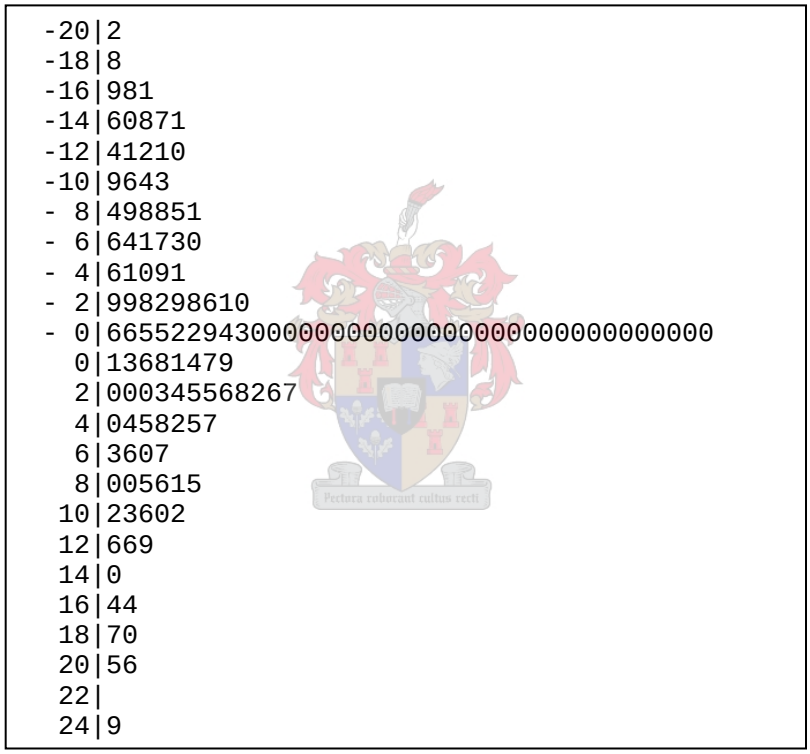


FIGURE 2

STEM-AND-LEAF PLOT OF STANDARDIZED RESIDUALS FOR SAMPLE A

The Q-plot displayed in Figure 3, provides an additional graphical display of residuals for Sample A by plotting the standardized residuals (horizontal axis) against the quantiles of the normal distribution (Diamantopoulos & Siguaw, 2000). The Q-plot indicates good model fit as there is a relatively small angular deviation of the standardized residuals for all pairs of observed variables from the 45° reference line in the Q-plot, especially in the upper and lower regions of the X-axis.

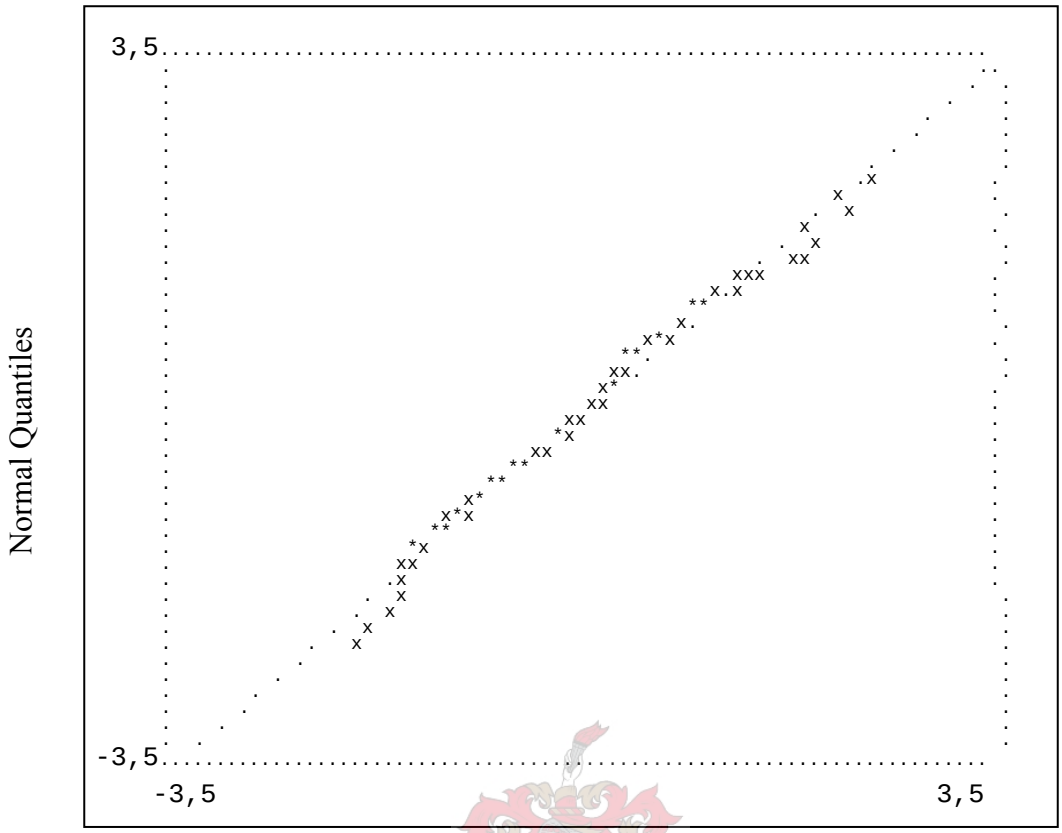


FIGURE 3

Q-PLOT OF STANDARDIZED RESIDUALS FOR SAMPLE A

4.4.3 Model modification indices for Sample A

Given the results presented thus far, the model depicted in equation 1 seems to fit closely to the data of Sample A and does not appear to indicate the need to consider the addition of one or more paths in order to improve the fit of the model. Examining the modification indices calculated for the currently fixed parameters of the model provides an additional way of determining if one or more paths would significantly improve the fit of the model. Model modification indices calculated by LISREL serve to estimate the decrease that should occur in the χ^2 statistic if parameters that are currently fixed are set free and the model re-estimated. Modification index with large values (> 6,6349) identify currently fixed parameters that would improve the fit of the model significantly if set free ($p < 0,01$) (Diamantopoulos & Siguaw, 2000). Any alteration to the model as suggested by modification indices should, however, only be considered if such alterations are substantively justifiable (Diamantopoulos & Siguaw, 2000;

Kelloway, 1998). In other words such alterations should make sense in relation to the theory of unit performance that underpins the PI.

The modification indices calculated for the Λ_X matrix identify nine additional paths that would significantly improve the fit of the PI measurement model for Sample A if they were set free. The largest modification indices suggests freeing the path from the Climate to Grow1 (15,40), from Core People Processes to Grow1 (13,59), and from Future Growth to Mark2 with large completely standardized expected change values for χ^2 (0,69; 0,74, -1,05). Although these modification indices suggest that substantial improvement in fit might be obtained from making one or more of these modifications, it is not possible to construct a theoretical justification for making any post hoc modifications to the measurement model based on this information. The significant modification indices suggest that one or more of the items included in the item parcels also systematically reflect one or more other latent variables than those they were designed to reflect. The primary question of interest is to what extent item parcels succeeded in reflecting the latent variable they were designated to reflect. The question whether item parcels also (assuming satisfactory high completely standardized factor loadings) systematically other latent variables seem of secondary interest. If a strategic decision would be taken to use complex items in the PI a prudent option would be to construct items with the explicit intent to reflect more than one latent variable and to fit a measurement model that reflects that design intention. It would after all be pointless to allow for cross-loading in the measurement model when this is not reflected in the scoring key of the PI.

Furthermore, the modification indices calculated for the Λ_X matrix suggest freeing the path from the sub-scale Production and Efficiency to both Employee Satisfaction item parcels, Satis1 (7,03) and Satis2 (6,93), as well as from Capacity to Satis1 (8,50). The magnitude of the modification index values taken in conjunction with the magnitude and sign of the standardized expected change values with completely standardized expected change values for χ^2 of 0,19, -0,18, and -0,29 respectively in this case, however, argue less convincingly for the freeing of the paths in question.

For the purpose of this study a conservative approach of upholding the original design intentions of the PI will be followed and consequently no allowance will be made for cross-loading of item parcels even though it could significantly improve the fit of the measurement model. The relative

small number of large modification index values calculated for the Λ_X matrix moreover supports the previous relatively optimistic conclusion on model fit.

4.4.4 Assessment of the first-order factor model

The completely standardized factor loading matrix (Λ_X) displayed in Table 19 reflects the regression of X_j on ξ_i and is used to evaluate the significance of the first-order factor loadings hypothesized by the proposed measurement model expressed as equation 1. The completely standardized λ parameter estimates reflect the average change in standard deviation units in a manifest variable X directly resulting from a one standard deviation change in a first-order exogenous latent variable ξ to which it has been linked, holding the effect of all other variables constant. The results depicted in Table 19 indicate that all proposed first-order factor loadings are significant ($p<0,05$). The fit of the model should therefore deteriorate significantly if any of the existing paths in the measurement model would be pruned away by fixing the corresponding parameters in Λ_X to zero and thus effectively eliminating the item parcels in question from their current sub-scales (Kelloway, 1998). None of the existing paths in the model thus appear to be redundant and all item parcels thus significantly reflect the unit performance dimension it was designed to denote. Moreover, Table 19 shows that the indicator variables generally load quite high on the first-order factors to which they have been linked. The two item parcels Prod1 and Grow2 should, however, be singled out as two exceptions to the rule.

TABLE 19

COMPLETELY STANDARDIZED FACTOR LOADING MATRIX (Λ_X) FOR SAMPLE A

Product		Core people		Climate		Satisfaction	
	0,76		0,86		0,93		0,80
Prod1	(0,04)	Core1	(0,03)	Clim1	(0,03)	Satis1	(0,04)
	12,53*		18,04*		20,19*		15,58*
	0,87		0,86		0,88		0,79
Prod2	(0,04)	Core2	(0,03)	Clim2	(0,04)	Satis2	(0,04)
	15,47*		17,61*		18,50*		15,23*
Adaptability		**Capacity**		**Market standing**		**Future growth**	
	0,90		0,85		0,79		0,88
Adapt1	(0,03)	Capac1	(0,03)	Mark1	(0,04)	Grow1	(0,04)
	18,93*		16,83*		14,14*		15,03*
Adapt2	0,85		0,80		0,78		0,65
	(0,04)	Capac2	(0,04)	Mark2	(0,04)	Grow2	(0,04)
	17,03*		15,11*		14,62*		10,95*

* t-values >|1,96| indicate significant path coefficients; values in brackets represent standard error estimates

Examining the variance explained in the item parcels reveals a similar picture (see also Table 17, page 46). The total variance in the i^{th} item parcel (X_i) can be decomposed into variance due to variance in the latent variable the item parcel was designed to reflect (ξ_j), variance due to variance in other systematic latent effects the item parcel was not designed to reflect and random measurement error. The latter two sources of variance in the item parcel are acknowledged in equation 1 through the measurement error term δ_i. The measurement error terms δ thus does not differentiate between systematic and random sources of error or non-relevant variance. The square of the completely standardized factor loadings λ_{ij} given in Table 19 could be interpreted as the proportion systematic-relevant item parcel variance given that each item parcels loads on one latent variable only. Since reliability could be defined as the extent to which variance in item parcels can be attributed to systematic sources, irrespective of whether the source of variance is relevant to the measurement intention or not, the completely standardized factor loading values shown in Table 19 could, therefore, be simultaneously interpreted as lower bound estimates of the item reliabilities (Diamantopoulos & Siguaw, 2000; Jöreskog & Sörbom, 1996a). Thus, given the modification indices calculated for Λ_X shown in Table 19 the extent to which the true item reliabilities would be under-estimated is not likely to be considerable and in most cases the item parcels seem to provide relatively uncontaminated reflections of their designated latent dimensions.

The proportion of item parcel variance that is explained by the latent variable it has been designated to reflect in terms of the measurement model (i.e. equation 1) is indicated by the squared multiple correlations for the observed indicator variables as shown in Table 20. Table 20 again reveals that the success with which Prod1 and Grow2 provide operational measures of the respective latent unit performance dimensions they are meant to reflect is not quite satisfactory. These item parcels are providing relatively contaminated reflections of their designated latent dimension. In the case of the latter item parcel this may be due to a problem for quit a number of units to practically apply the concept of future growth to their unit.

TABLE 20

SQUARED MULTIPLE CORRELATIONS FOR ITEM PARCELS FOR SAMPLE A

Prod1	**Prod2**	**Core1**	**Core2**	**Clim1**	**Clim2**	**Satis1**	**Satis2**
0,57	0,75	0,74	0,73	0,86	0,77	0,65	0,63
Adapt1	**Adapt2**	**Capac1**	**Capac2**	**Mark1**	**Mark2**	**Grow1**	**Grow2**
0,82	0,72	0,71	0,64	0,62	0,60	0,78	0,43

The phi-matrix of correlations between the eight latent unit performance sub-scales is shown in Table 21. The off-diagonal elements of the Φ-matrix are the sub-scale correlations disattenuated for measurement error. As the Φ-matrix is positive definite and off-diagonal entries do not exceed unity, the results tend to provide some support for the discriminant validity of the first-order factors. All of the 28 correlations are significant ($p<0,01$) although only 5 correlations are highly significant.

TABLE 21

COMPLETELY STANDARDIZED PHI (Φ) MATRIX FOR SAMPLE A

	Production & Efficiency	Core people	Climate	Satisfaction	Adaptability	Capacity	Market Standing	Future Growth
Product & Efficiency	1,00							
Core People	0,67	1,00						
Climate	0,60	0,83	1,00					
Satisfaction	0,63	0,89	0,86	1,00				
Adaptability	0,53	0,76	0,68	0,82	1,00			
Capacity	0,50	0,66	0,62	0,70	0,83	1,00		
Market Standing	0,60	0,71	0,67	0,75	0,82	0,88	1,00	
Future Growth	0,54	0,71	0,62	0,66	0,77	0,87	0,87	1,00

These correlations are to a certain extent expected given the nature of the underlying unit performance model and the results obtained by Henning *et al. (*2003) that provides support that specific relationships may exist between latent variable as displayed in the PI structural model (Figure 1, see page 8). As such, the nature of the phi matrix may be seen as an expression of the complexity of unit performance in the sense that the various sub-scales comprising unit performance may well directly or indirectly causally influence each other. In particular attention may be drawn to the high correlations among two sets of latent variables, one set including Employee Satisfaction, Core People Processes and Climate, and the other set including Market Standing, Capacity and Future Growth. It does not seem altogether unreasonable to argue that high correlations may well exist between these sets of latent variables as, for example, Core

People Processes may logically be related, if not a key determining factor, of Climate and Employee Satisfaction in a work unit. Similarly, it may be suggested that it is logical to expect relationships among Capacity, Market Standing and Future Growth. In summary, the above results suggest that the indicator variables do generally succeed in providing empirical grasp on the underlying latent variables they were meant to reflect.

4.4.5 Summary of model fit assessment for Sample A

The model fit statistics for Sample A seem to unanimously suggest that the model fits closely to the data of Sample A, and that the model fully captures the true complexity of the underlying PI model without the need for additional paths as it outperforms the independence and saturated models. Analysis of residuals supported this conclusion as the standardized residuals were distributed approximately symmetrical around a median standardized residual of zero, no large positive and negative residuals were present, and the standardized residuals for all pairs of observed variables demonstrated only a small angular deviation of from the 45° reference line in the Q-plot. These observations support the conclusion that the observed covariance terms in the observed sample covariance matrix are estimated reasonably well by the derived model parameter estimates. Model modification indices calculated for the Λ_X matrix identified nine additional paths that would significantly improve the fit of the PI measurement model for Sample A if they were set free. The latter finding tends to temper the optimistic model fit conclusion to some extent although not overly so given the limited number of large values.

Analysis of the completely standardized factor loading matrix and squared multiple correlations for Λ_X indicated that in most cases the item parcels seem to provide relatively uncontaminated reflections of their designated latent dimensions, although there was some indication that Prod1 and Grow2 could be providing relatively contaminated reflections of their designated latent dimension. Lastly, the phi matrix identified two sets of sub-scales that are more highly correlated than the rest of the unit performance dimensions. These correlations might be an expression of the direct and indirect causal influences existing between these latent variables. Relationships between these variables do not, seem altogether unreasonable and are in fact hypothesized to exist by the basic PI structural model (Henning *et al.*, 2003). In sum, when the abovementioned results are considered in conjunction the measurement model depicted in equation 1 seems to closely fit the data of Sample A.

4.5 RESULTS FOR SAMPLE B

4.5.1 Overall fit assessment for Sample B

The full spectrum of indices provided by LISREL to assess the absolute and comparative fit of the proposed measurement model with the data from Sample B is presented in Table 22. An admissible final solution of parameter estimates for the PI measurement model was obtained after 9 iterations.

TABLE 22

GOODNESS-OF-FIT INDICATORS FOR SAMPLE B

Degrees of Freedom = 76
Minimum Fit Function Chi-Square = 131,75 (P = 0,00)
Normal Theory Weighted Least Squares Chi-Square = 129,71 (P = 0,00012)
Satorra-Bentler Scaled Chi-Square = 115,77 (P = 0,0022)
Chi-Square Corrected for Non-Normality = 142,12 (P = 0,00)
Estimated Non-centrality Parameter (NCP) = 39,77
90 Percent Confidence Interval for NCP = (14,67 ; 72,82)

Minimum Fit Function Value = 0,35
Population Discrepancy Function Value (F0) = 0,11
90 Percent Confidence Interval for F0 = (0,039 ; 0,19)
Root Mean Square Error of Approximation (RMSEA) = 0,037
90 Percent Confidence Interval for RMSEA = (0,023 ; 0,051)
P-Value for Test of Close Fit (RMSEA < 0,05) = 0,94

Expected Cross-Validation Index (ECVI) = 0,63
90 Percent Confidence Interval for ECVI = (0,56 ; 0,72)
ECVI for Saturated Model = 0,73
ECVI for Independence Model = 25,91

Chi-Square for Independence Model with 120 Degrees of Freedom = 9658,77
Independence AIC = 9690,77
Model AIC = 235,77
Saturated AIC = 272,00
Independence CAIC = 9769,60
Model CAIC = 531,38
Saturated CAIC = 942,06

Normed Fit Index (NFI) = 0,99
Non-Normed Fit Index (NNFI) = 0,99
Parsimony Normed Fit Index (PNFI) = 0,62
Comparative Fit Index (CFI) = 0,99
Incremental Fit Index (IFI) = 0,99
Relative Fit Index (RFI) = 0,98

Critical N (CN) = 306,40

Root Mean Square Residual (RMR) = 0,013
Standardized RMR = 0,028
Goodness-of-fit Index (GFI) = 0,96
Adjusted Goodness-of-fit Index (AGFI) = 0,93
Parsimony Goodness-of-fit Index (PGFI) = 0,54

The Satorra-Bentler χ^2 test statistic (115,77) is significant ($p<0,01$) thus resulting in a rejection of the null hypothesis of exact model fit (H_0: $\Sigma=\Sigma(\theta)$). The normed χ^2 (1,52) indicates that the measurement model is neither 'over-fitted' or 'under-fitted' but rather demonstrates acceptable fit to the data. The estimated λ value for Sample B (39,77) is not very high. The 90 percent confidence interval for NCP (14,67 ; 72,82) also ranges across acceptable values. This implies good model fit as the estimated discrepancy between the observed (Σ_0) and estimated population covariance ($\widetilde{\Sigma}_0$) matrices is not very high (Diamantopoulos & Siguaw, 2000). The RMSEA value (0,037) for Sample B easily meets the criteria of close model fit (Browne & Cudeck, 1993). The 90 percent confidence interval for RMSEA (0,023 ; 0,051) also indicates good fit. Similarly, the test of close fit performed by LISREL shows that the conditional probability of the obtained RMSEA value of 0,037 under H_0: RMSEA$\leq$0,05 is sufficiently large (0,94) not to reject the close fit null hypothesis. These results suggest that close model fit has been achieved for Sample B.

The model ECVI (0,63) also provides support for good model fit as it is far smaller than both the value obtained for the independence model (25,91) and the ECVI value associated with the saturated model (0,73), and thus most likely to be replicated in a cross-validation sample (Diamantopoulos & Siguaw, 2000). The parsimonious normed fit index (PNFI = 0,62) and the parsimonious goodness-of-fit index (PGFI = 0,54) approach model fit from this perspective, although, as mentioned above, these indices do not provide valuable information for the purpose of this analysis. On the other hand, the model AIC (235,77) suggests that the fitted measurement model provides a more parsimonious fit than the independent/null model (9690,77) and the saturated model (272,00) (Kelloway, 1998). Likewise, the model CAIC (531,38) indicates that the fitted measurement model provides a more parsimonious fit than both the independent/null model (9769,60) and the saturated model (942,06). These results imply that the measurement model is indeed the most parsimonious and does not require additional paths.

The indices of relative fit given in Table 22 all exceed the critical value of 0,90 and therefore indicate good comparative fit when compared to the independence model (Diamantopoulos & Siguaw, 2000; Kelloway, 1998). These indices include the NFI (0,99), NNFI (0,99), CFI (0,99), IFI (0,99), and the RFI (0,98). The estimated CN value (306,4) is substantially above the recommended threshold value of 200. This implies that the model provides an adequate representation of the data (Diamantopoulos & Siguaw, 2000). In a similar vein, the RMR (0,013) and standardized RMR (0,028) indicate good fit. For Sample B, AGFI (0,93) and the GFI (0,96) both exceed 0,9 which indicates the model comes close to perfectly reproducing the sample

covariance matrix and therefore suggests good model fit (Jöreskog & Sörbom, 1993; Kelloway, 1998).

In summary, the above model fit statistics seems to indicate unambiguously that the model fits the data of Sample B closely. The model outperforms the independence model and the saturated model repeatedly which also implies that it fully captures the true complexity of the underlying PI model without the need for additional paths.

4.5.2 Examination of residuals

The stem-and-leaf plot of the standardized residuals for Sample B is provided in Figure 4.

```
- 4|3
- 3|
- 3|0
- 2|88
- 2|421
- 1|887776666665
- 1|4442
- 0|8877776666666665
- 0|4433322211111100000000000000000000000000000000000000000000
  0|1111222233333344
  0|5677777889
  1|01222444
  1|568
  2|
  2|79
  3|
  3|6
  4|
  4|
  5|4
```

FIGURE 4

STEM-AND-LEAF PLOT OF STANDARDIZED RESIDUALS FOR SAMPLE B

The distribution of standardized residuals appears to be distributed approximately symmetrical around a median standardized residual of zero. The average standardized residual reported earlier (0,028) also suggested that on average the elements of the observed sample covariance matrix are accurately reproduced by the parameter estimates of the model. However, the smallest (-4,30) and largest (5,40) standardized residuals fall well outside the 0,01 significance limits of ±2,58 which suggests that a number of the observed covariance terms in the observed sample covariance matrix are not estimated as well as they could be by the derived model parameter estimates. This goes

against the relatively positive conclusion of model fit that was suggested by the fit statistics earlier. The number of large and small standardized residuals are, however small. Only four (2,94%) large negative residual occur and only four (2,94%) large positive residuals.

The distribution of standardized residuals around the 45° reference line on a Q-plot indicates departure from normality and/or specification errors in a model. The Q-plot of standardized residuals for Sample B is presented as Figure 5. The obvious nonlinear pattern and deviation of the standardized residuals from the 45° reference line both in the upper and lower regions of the X-axis on the Q-plot suggests that the model does not fit the empirical data adequately. Specification errors are also indicated by the presence of outliers on the Q-plot (Diamantopoulos & Siguaw, 2000).

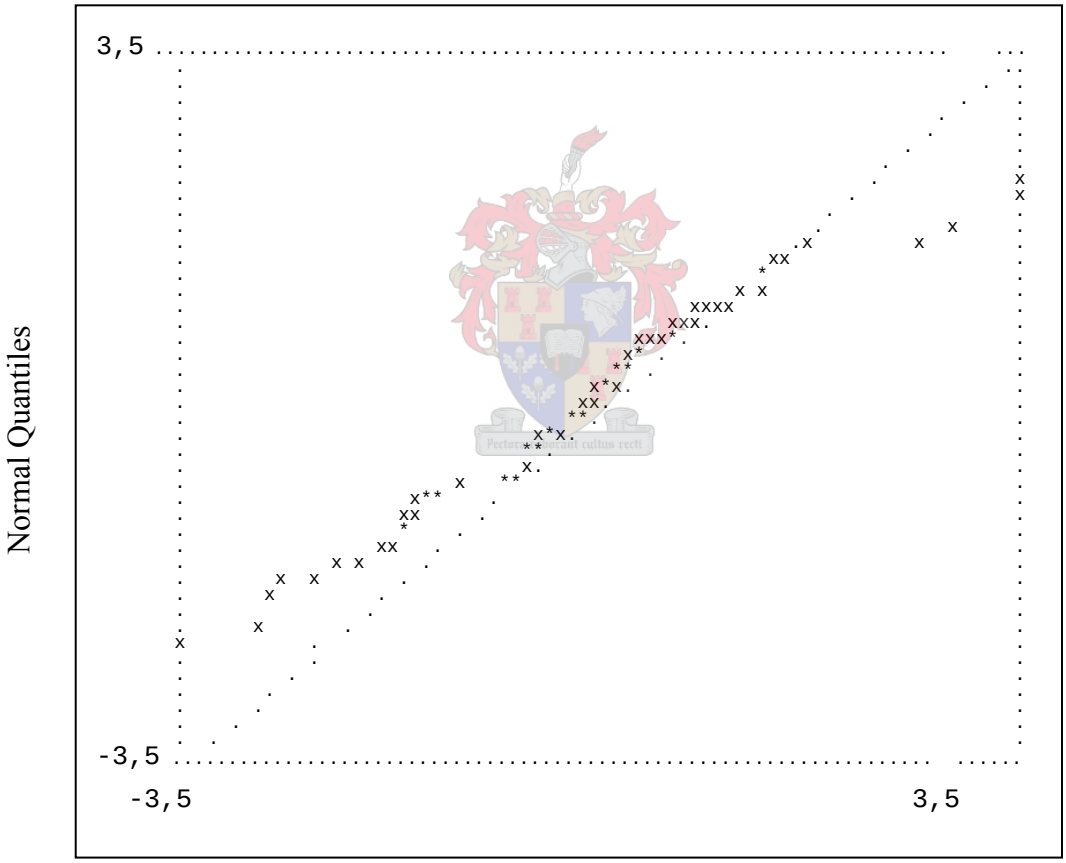


FIGURE 5

Q-PLOT OF STANDARDIZED RESIDUALS FOR SAMPLE B

Thus, although the abovementioned model fit statistics seem to unanimously suggest that the model fits closely to the data of Sample B and appears to capture the true complexity of the underlying PI model without the need for additional paths, the foregoing analysis of the

standardized residuals suggests that the addition of one or more paths may be required to improve the fit of the model. Examining the modification indices calculated for the currently fixed parameters of the model provides an additional way of determining if the addition of one or more paths would significantly improve the parsimonious fit of the model for Sample B.

4.5.3 Model modification indices for Sample B

The modification indices calculated for the Λ_X matrix identify four additional paths that, if set free, would significantly improve the fit of the PI measurement model for Sample B. The modification indices suggest freeing the path from the Climate to Grow2 (9,67), from Employee Satisfaction to Grow2 (8,18), from Future Growth to Mark1 (8,43) and Mark2 (7,87) with completely standardized expected change values for χ^2 (-0,22; -0,27; 0,25; -0,26). Although these modification indices suggest that some improvement in fit might be obtained from making one or more of these modifications, they are not very strong and the expected changes in χ^2 are not very large. Furthermore, freeing these specific elements of Λ_X does not appear to be easily justified. The same conservative argument put forward in the case of Sample A will again be used here to justify maintaining the status quo.

4.5.4 Assessment of the first-order factor model for Sample B

The completely standardized factor loading matrix (Λ_X) (Table 23) reflecting the regression of X_j on ξ_i, is used to evaluate the significance of the first-order factor loadings hypothesized by the proposed measurement model expressed as equation 1. The results depicted in Table 23 indicate that all proposed first-order factor loadings are significant ($p<0,05$). This means that none of the existing paths in the model appear to be redundant and all item parcels appear to significantly reflect the unit performance dimension they were designed to represent.

The proportion of item parcel variance that is explained by the latent variable it has been designated to reflect in terms of the measurement model (i.e., equation 1) is indicated by the squared multiple correlations for the observed indicator variables shown in Table 24. In most cases the item parcels seem to provide relatively uncontaminated reflections of their designated latent dimensions. Of concern, however, are the values for Mark1, Grow2 and Satis2 as they suggest a large proportion of variance in these item parcels cannot be explained by the latent variable it is designed to represent. These findings therefore suggest that these items might not

be satisfactory operational measures of their respective latent unit performance dimensions. Although the item parcels that were flagged as problematic in the two analyses, the R^2 pattern presented in Table 20 for sample A nonetheless agrees quite strongly (r=0,866) with the pattern presented in Table 24 for sample B.

TABLE 23

COMPLETELY STANDARDIZED FACTOR LOADING MATRIX (Λ_X)

Product		Core people		Climate		Satisfaction	
	0,80		0,88		0,93		0,80
Prod1	(0,03)	Core1	(0,03)	Clim1	(0,03)	Satis1	(0,03)
	16,97*		21,91*		23,52		17,65*
	0,87		0,82		0,86		0,73
Prod2	(0,03)	Core2	(0,03)	Clim2	(0,03)	Satis2	(0,04)
	19,09*		19,28*		22,31*		14,46*
Adaptability		**Capacity**		**Market standing**		**Future growth**	
	0,86		0,83		0,68		0,89
Adapt1	(0,03)	Capac1	(0,03)	Mark1	(0,04)	Grow1	(0,03)
	20,93*		17,50*		12,57*		17,69*
Adapt2	0,86		0,82		0,76		0,68
	(0,03)	Capac2	(0,03)	Mark2	(0,03)	Grow2	(0,04)
	20,75*		18,10*		16,55*		11,99*

* t-values >|1,96| indicate significant path coefficients; values in brackets represent standard error estimates

TABLE 24

SQUARED MULTIPLE CORRELATIONS FOR ITEM PARCELS FOR SAMPLE B

Prod1	**Prod2**	**Core1**	**Core2**	**Clim1**	**Clim2**	**Satis1**	**Satis2**
0,64	0,76	0,78	0,67	0,86	0,75	0,64	0,53
Adapt1	**Adapt2**	**Capac1**	**Capac2**	**Mark1**	**Mark2**	**Grow1**	**Grow2**
0,74	0,75	0,70	0,67	0,47	0,58	0,78	0,47

The phi-matrix of correlations between the eight latent unit performance sub-scales is shown in Table 25. As the Φ-matrix is positive definite and off-diagonal entries do not exceed unity, the results tend to provide some support for the discriminant validity of the first-order factors. All of the 28 correlations are significant (p<0,01) although only 3 correlations are highly significant. As mentioned above, correlations are to a certain extent expected given the nature of the underlying unit performance model and the results obtained by Henning *et al.* (2002) that provides support that specific relationships may exist between latent variable as displayed in the PI structural model (Figure 1, see page 8). Comparable to Sample A, results for Sample B indicate strong relationships may exist between Employee Satisfaction, Core People Processes and Climate, but to a somewhat lesser extent between Market standing, Capacity and Future Growth. The above results suggest that the indicator variables, for the most part at least, succeed in providing empirical grasp on the underlying latent variables they were meant to reflect.

TABLE 25

COMPLETELY STANDARDIZED PHI (Φ) MATRIX FOR SAMPLE B

	Production & Efficiency	Core people	Climate	Satisfaction	Adaptability	Capacity	Market Standing	Future Growth
Product & Efficiency	1,00							
Core People	0,68	1,00						
Climate	0,56	0,87	1,00					
Satisfaction	0,64	0,87	0,89	1,00				
Adaptability	0,54	0,75	0,70	0,75	1,00			
Capacity	0,43	0,60	0,55	0,64	0,69	1,00		
Market Standing	0,60	0,70	0,65	0,74	0,74	0,78	1,00	
Future Growth	0,50	0,56	0,50	0,59	0,47	0,54	0,68	1,00

4.5.5 Summary of model fit assessment for Sample B

As with Sample A, the model fit statistics seems to indicate unambiguously that the model fits closely to the data of Sample B and fully captures the true complexity of the underlying PI model without the need for additional paths as the model outperforms the independence and saturated models repeatedly. The stem-and-leaf plot and the Q-plot of the standardized residuals for Sample B did not, however, support this positive conclusion. Rather, a number of large positive and negative residuals on the stem-and-leaf plot as well as outliers and a decidedly non-linear patter and deviation of the standardized residuals from the 45° reference line both in the upper and lower regions of the X-axis on the Q-plot, suggested that at least eight observed covariance terms in the observed sample covariance matrix are not estimated as well as they could be by the derived model parameter estimates. Likewise, the analyses of the modification indices calculated for the currently fixed parameters of the model also suggested that the addition of one or more paths would significantly improve the parsimonious fit of the model for Sample B. However, several reasons were presented against freeing these specific elements of $\Lambda_{X,}$ including the fact that this could not be theoretically justified without further research, and the fact that the modification indices and the expected changes values for λ_{ij} were not very large.

Analysis of the completely standardized factor loading matrix and squared multiple correlations for Λ_X indicated that in most cases the item parcels seem to provide relatively uncontaminated reflections of their designated latent dimensions, although there was some indication that Mark1, Grow2 and (to a lesser extent) Satis2 might not be satisfactory operational measures of their respective latent unit performance dimension. The phi matrix identified three highly correlated sub-scales that may directly or indirectly causally influence each other. Relationships among these three variables could be justified and presented no reason for concern. In sum, considering the abovementioned results in an integrated manner seems to support the conclusion that the measurement model depicted in equation 1 seems to fit closely to the data of Sample B.

The finding that the same measurement model fits the data of Sample A and Sample B closely implies that the PI measurement model depicted as equation 1 demonstrates configural invariance (Vandenberg & Lance, 2000). The number of latent variables and the same pattern of factor loadings are required to explain the observed covariance matrices in samples A and B. To determine whether the magnitude of the measurement model parameter estimates significantly differ across the model fitted to Sample A and the model fitted to Sample B requires the establishment of a set of baseline multi-group model fit indices.

4.6 EVALUATION OF THE UNCONSTRAINED MULTI-GROUP MEASUREMENT MODEL

At least reasonable model fit is required of both Sample A and Sample B to justify cross-validating the measurement model across these two samples. For this reason the measurement model was fitted to both samples independently. As the results for both sample A and B indicated close model fit, there is sufficient reason to investigate whether the parameter estimates can be considered equal across samples. As a first step in this cross-validation process it is necessary to first describe the degree of measurement model fit when the measurement model is fitted to Sample A and Sample B simultaneously in a multi-group analysis with no parameters constrained. The resultant global fit statistics will be used as a baseline to evaluate subsequent restrictions imposed on the model. The following section presents a summary of the fit statistics for the model with no parameters constrained.

4.6.1 Model Identification for the multi-group model with no parameters constrained

The question of model identification had to be examined prior to confronting the model with data to determine if the model and the data would permit the determination of unique estimates for the freed parameters in the model when used for cross-validation analyses on the two samples combined (Diamantopoulos & Siguaw, 2000). The measurement model depicted as equation 1 satisfies the two requirements suggested by Diamantopoulos and Siguaw (2000) and MacCallum (1995). A definite scale is established for each latent variable as each latent variable is treated as a (0; 1) standardized variable. In addition, the number of model parameters to be estimated (t=120) do not exceed the collective number of unique covariance terms for the observed sample covariance matrices for both Sample A and Sample B (272) (Diamantopoulos & Siguaw, 2000; MacCallum, 1995). The degrees of freedom thus are 152.

4.6.2 Goodness-of-fit of the multi-group measurement model with no parameters constrained

The goodness-of-fit indices provided by LISREL to assess the absolute and comparative fit of the measurement model fitted to both samples simultaneously with no parameters constrained is presented in Table 26. The question is how well the two sets of parameter estimates derived freely for the same model from the data of the two samples succeed in reproducing/explaining the observed covariance matrix. An admissible final solution of parameter estimates for the PI measurement model was obtained after 14 iterations.

The Satorra-Bentler χ^2 test statistic (256,77) is significant ($p<0,01$) thus resulting in a rejection of the null hypothesis of exact model fit (H_0: $\Sigma=\Sigma(\theta)$). In a multi-group analysis, the chi-square is a measure of fit of the model across all groups and cannot be decomposed into a chi-square for each group separately (Du Toit *et al.*, 2001). This perspective also applies to the other fit indices. The normed χ^2 (1,69) indicates that the measurement model is neither 'over-fitted' or 'under-fitted' but rather demonstrates acceptable fit to the data. The estimated λ value (104,77) and the 90 percent confidence interval for NCP (64,44; 152,99) imply good model fit, as the estimated discrepancy between the observed and estimated population covariance matrices is not very high (Diamantopoulos & Siguaw, 2000).

TABLE 26

GOODNESS-OF-FIT INDICATORS FOR THE MULTI-GROUP MEASUREMENT MODEL WITH NO PARAMETERS CONSTRAINED

Degrees of Freedom = 152
Minimum Fit Function Chi-Square = 273,17 (P = 0,00)
Normal Theory Weighted Least Squares Chi-Square = 269,75 (P = 0,00)
Satorra-Bentler Scaled Chi-Square = 256,77 (P = 0,00)
Chi-Square Corrected for Non-Normality = 416,13 (P = 0,0)
Estimated Non-centrality Parameter (NCP) = 104,77
90 Percent Confidence Interval for NCP = (64,44 ; 152,99)

Minimum Fit Function Value = 0,42
Population Discrepancy Function Value (F0) = 0,16
90 Percent Confidence Interval for F0 = (0,099 ; 0,24)
Root Mean Square Error of Approximation (RMSEA) = 0,046
90 Percent Confidence Interval for RMSEA = (0,036 ; 0,056)
P-Value for Test of Close Fit (RMSEA < 0,05) = 1,00

Expected Cross-Validation Index (ECVI) = 0,76
90 Percent Confidence Interval for ECVI = (0,70 ; 0,84)
ECVI for Saturated Model = 0,42
ECVI for Independence Model = 28,33

Chi-Square for Independence Model with 240 Degrees of Freedom = 18382,70
Independence AIC = 18446,70
Model AIC = 496,77
Saturated AIC = 544,00
Independence CAIC = 18622,06
Model CAIC = 1154,38
Saturated CAIC = 2034,57

Normed Fit Index (NFI) = 0,99
Non-Normed Fit Index (NNFI) = 0,99
Parsimony Normed Fit Index (PNFI) = 0,62
Comparative Fit Index (CFI) = 0,99
Incremental Fit Index (IFI) = 0,99
Relative Fit Index (RFI) = 0,98

Critical N (CN) = 466,13

Contribution to Chi-Square = 141,42
Percentage Contribution to Chi-Square = 51,77
Root Mean Square Residual (RMR) = 0,015
Standardized RMR = 0,029
Goodness-of-fit Index (GFI) = 0,94

The RMSEA value (0,046) meets the ≤0,05 criterion of close model fit (Browne & Cudeck, 1993). The 90 percent confidence interval for RMSEA (0,036 ; 0,056) indicates good fit. Similarly, the test of close fit performed by LISREL shows that the conditional probability of obtaining the observed sample RMSEA value of 0,046 under H0: population RMSEA (0,05 is 1,00, which implies that the null hypothesis of close model fit cannot be rejected. The model ECVI (0,76) is far smaller than both the value obtained for the independence model (28,33) but larger than the ECVI value associated with the saturated model (0,42). This suggests that a

model more closely resembling the saturated model may have a better chance of being replicated in a cross-validation multi-sample analysis than the fitted model (Diamantopoulos & Siguaw, 2000). On the other hand, the model AIC (496,77) suggests that the fitted measurement model provides a more parsimonious fit than the independent/null model (18446,77) and the saturated model (544,00) (Kelloway, 1998). Likewise, the model CAIC (1154,38) indicates that the fitted measurement model provides a more parsimonious fit than both the independent/null model (18622,06) and the saturated model (2034,57). In contrast to model ECVI, the model AIC and CAIC results imply that the measurement model is indeed the most parsimonious and does not require additional paths.

The indices of relative fit given in Table 26 all exceed the critical value of 0,90 and therefore indicate good comparative fit when compared to the independence model (Diamantopoulos & Siguaw, 2000; Kelloway, 1998). These indices include the NFI (0,99), NNFI (0,99), CFI (0,99), IFI (0,99), and the RFI (0,98). The estimated CN value (466,13) is substantially above the recommended threshold value of 200, which indicates that the model provides an adequate representation of the data (Diamantopoulos & Siguaw, 2000). In a similar vein, the RMR (0,015) and standardized RMR (0,029) suggest good fit. The GFI (0,94) exceeds 0,9, which indicates the model comes close to perfectly reproducing the sample covariance matrix and therefore suggests good model fit (Jöreskog & Sörbom, 1993; Kelloway, 1998).

In summary, almost all of the above model fit statistics suggest that the measurement model with unconstrained parameters fits the data of Sample A and Sample B closely. The question subsequently arises whether the model parameters freely estimated from the data of the two samples differ significantly across the two samples. The model might fit the data of both samples closely but the parameter estimates might nonetheless differ markedly across the two samples even though they represent the same target population. If so, confidence in the original claims made by Spangenberg and Theron (2004) and Henning *et al.* (2003) would be seriously eroded. Consequently these results necessitate the further examination of the extent to which the measurement model cross-validates successfully across these two samples.

4.7 MEASUREMENT INVARIANCE TESTS

4.7.1 Omnibus test: parameters set to be equal

Cross-validation of the measurement model refers to the ability of the model to be invariant across two or more random samples from the same population (Mels, 2003) and may be determined by investigating the stability of the model parameter estimates when the model is fitted to two samples simultaneously from the same population. The omnibus test of measurement invariance tests the null hypothesis ($\Sigma^g = \Sigma^{g'}$) that the model fits closely to the data across both samples simultaneously when all parameter estimates are set to be equal across samples (Vandenberg & Lance, 2000). In effect the omnibus test compares the model fit obtained when the model is fitted to the data of two samples from the same population simultaneously with the condition that all parameter estimates are set to be equal across samples, with the model fit when the model is fitted with no parameters constrained. Specifically the question is whether imposing the equality constraint on the model parameter estimates results in a significant deterioration in the model fit. If a significant increase in the Satorra-Bentler chi-square does not result from the imposition of the equality constraint the foregoing null hypothesis will not be rejected. Failure to reject the null hypothesis means the PI may be considered measurement invariant across the samples and subsequent tests of measurement invariance are not required. Rejection of the null hypothesis suggests that the model does not cross-validate well across different samples (Vandenberg & Lance, 2000).

The decision of whether or not to reject the null hypothesis is based on the significance of the chi-square statistic and other overall goodness-of-fit indices. The results for the omnibus test are given in Table 27. The Satorra-Bentler χ^2 test statistic (358,04) is significant ($p<0,01$) which means that the null hypothesis of exact model fit should be rejected (Vandenberg & Lance, 2000). In contrast, the other fit indices almost unanimously indicate the model fits closely when fitted to the data from Sample A and Sample B simultaneously with all parameter estimates constrained to be equal. These descriptive indices are summarised below.

The normed χ^2 (1,69) indicates that the measurement model is neither 'over-fitted' or 'under-fitted' but rather demonstrates acceptable fit to the data. The estimated λ value (146,04) and the 90 percent confidence interval for NCP (97,72; 202,25) suggest good model fit, as the estimated discrepancy between the observed matrices is not very high (Diamantopoulos & Siguaw, 2000). The RMSEA value (0,046) meets the $\leq$0,05 criteria of close model fit (Browne & Cudeck, 1993),

and the 90 percent confidence interval for RMSEA (0,038; 0,054) indicates good fit. The test of close fit performed by LISREL, moreover, shows that the conditional probability of the obtained RMSEA value of 0,037 under H0: RMSEA≤0,05 is sufficiently large (1,00) not to reject the close fit null hypothesis.

TABLE 27

GOODNESS-OF-FIT INDICATORS FOR OMNIBUS TEST

Degrees of Freedom = 212
Minimum Fit Function Chi-Square = 367,47 (P = 0,00)
Normal Theory Weighted Least Squares Chi-Square = 370,50 (P = 0,00)
Satorra-Bentler Scaled Chi-Square = 358,04 (P = 0,00)
Chi-Square Corrected for Non-Normality = 754,16 (P = 0,0)
Estimated Non-centrality Parameter (NCP) = 146,04
90 Percent Confidence Interval for NCP = (97,72 ; 202,25)
Minimum Fit Function Value = 0,57
Population Discrepancy Function Value (F0) = 0,22
90 Percent Confidence Interval for F0 = (0,15 ; 0,31)
Root Mean Square Error of Approximation (RMSEA) = 0,046
90 Percent Confidence Interval for RMSEA = (0,038 ; 0,054)
P-Value for Test of Close Fit (RMSEA < 0,05) = 1,00
Expected Cross-Validation Index (ECVI) = 0,74
90 Percent Confidence Interval for ECVI = (0,66 ; 0,82)
ECVI for Saturated Model = 0,42
ECVI for Independence Model = 28,33
Chi-Square for Independence Model with 240 Degrees of Freedom = 18382,70
Independence AIC = 18446,70
Model AIC = 478,04
Saturated AIC = 544,00
Independence CAIC = 18622,06
Model CAIC = 806,84
Saturated CAIC = 2034,57
Normed Fit Index (NFI) = 0,98
Non-Normed Fit Index (NNFI) = 0,99
Parsimony Normed Fit Index (PNFI) = 0,87
Comparative Fit Index (CFI) = 0,99
Incremental Fit Index (IFI) = 0,99
Relative Fit Index (RFI) = 0,98
Critical N (CN) = 465,89
Contribution to Chi-Square = 192,24
Percentage Contribution to Chi-Square = 52,31
Root Mean Square Residual (RMR) = 0,046
Standardized RMR = 0,098
Goodness-of-fit Index (GFI) = 0,92

The results indicate that the model ECVI (0,74) is far smaller than both the value obtained for the independence model (28,33) but larger than the ECVI value associated with the saturated model (0,42). This suggests that a model more closely resembling the saturated model may have

a better chance of being replicated in a cross-validation sample than the fitted model (Diamantopoulos & Siguaw, 2000). In comparison, the model AIC and CAIC results imply that the measurement model is indeed the most parsimonious and does not require additional paths as the model AIC (478,04) provides a more parsimonious fit than the independent model (18446,70) and saturated model (544,00) (Kelloway, 1998). Similarly, the model CAIC (808,84) indicates that the fitted measurement model provides a more parsimonious fit than both the independent model (18622,06) and the saturated model (2034,57). The indices of relative fit all exceed the critical value of 0,90 and therefore indicate good comparative fit when compared to the independence model, and the estimated CN value (466,13) suggests that the model provides an adequate representation of the data (Diamantopoulos & Siguaw, 2000). In a similar vein, the RMR (0,046) and standardized RMR (0,098) suggest good fit. The GFI (0,92) exceeds 0,9, which indicates the model comes close to perfectly reproducing the sample covariance matrix and therefore suggests good model fit (Jöreskog & Sörbom, 1993; Kelloway, 1998).

In summary, the descriptive fit indices, given in Table 27, suggest the measurement model with parameters constrained to be equal across Sample A and Sample B still fits closely. The critical question, however, is whether the fit deteriorated significantly when the equality constraint was imposed on the model in comparison to the multi-group analysis in which parameters estimates were allowed to differ across samples. If the model parameters do in fact differ across the two samples the multi-group fit will deteriorate when parameter estimates are constrained to be equal across samples. This can be evaluated by calculating the difference in the Satorra-Bentler χ^2 fit statistic achieved under the constrained and unconstrained conditions respectively. The χ^2 difference will itself follow a chi-square distribution with degrees of freedom equal to the difference in degrees of freedom between the constrained and unconstrained conditions (Mels, 2003). The results of this calculation are shown in Table 28.

TABLE 28

CHI-SQUARE DIFFERENCE TEST OF MEASUREMENT INVARIANCE

Hypothesis	Chi-square value	Degrees of freedom	Critical χ^2 (p=0,05)
Model with total invariance equality constraints imposed (H_0)	358,04	212	
Model with no parameters constrained (H_a)	256,77	152	
Difference	101,27**	60	79,8

(*= significant at the p<0,01 level; ** = significant at p< 0,05)

Table 28 reveals that the freeing of the equality constraint results in a significant improvement in the multiple-group model fit. This result is not altogether surprising, though, as in practical applications full measurement invariance frequently does not hold, especially in social science research where some differences between groups are to be expected (Steenkamp & Baumgartner, 1998). Moreover it needs to be considered whether the rejection of the measurement invariance hypothesis could not possibly be due to the fact that the samples were not truly two independent random samples from the same target population. Failure to reject the null hypothesis would have implied that the respondents of each sample employed the same conceptual frame of reference when completing the PI items and provide sufficient evidence of measurement invariance to justify other research that examines group differences in relation to the PI's underlying constructs (Vandenberg & Lance, 2000).

A limitation of the omnibus test is that it is does not provide information on the potential source of measurement invariance. As such, it is quite uninformative. Further tests of measurement invariance that test a series of increasingly restrictive hypothesis are, therefore, required to identify the source of the non-equivalence (Vandenberg & Lance, 2000).

4.7.2 Test of metric invariance (invariance of factor loadings)

Given that the test of full measurement invariance could not support the conclusion of measurement invariance across the samples, a stronger test of factorial invariance may be employed (Vandenberg, 2000). The test of metric invariance tests the null hypothesis ($\Lambda^{g}_{x} = \Lambda^{g}_{x}{}'$) that the factor loadings of item parcels on latent variables are equivalent across both samples. As such, tests of metric invariance have the null hypothesis that factor loadings for like items are invariant across groups (Vandenberg & Lance, 2000). This test may be operationalised by fitting a model in which the lambda-X matrices, in addition to the number of factors and the factor loading patterns, are constrained to be equal across samples. Failure to reject the null hypothesis will indicate that the factor loadings are invariant across both samples and that non-equivalence, as indicated by the results of the configural invariance test, can be attributed to other parameter estimates in the measurement model.

The Chi-square difference test is used to establish if a significant ($p<0,05$) difference exists between the Satorra-Bentler Chi-square values for the model with metric invariance constraints imposed and the model with no parameters constrained (Mels, 2003). If the fit does not

deteriorate significantly under the lambda-X equality constraint, equality of factor loadings can thus be assumed. Failure to reject the null hypothesis would imply that slope of the regression of the item parcels on the latent variables they are designed to represent are the same across the two samples. Parameter differences would, therefore, have to exist elsewhere in the model. Should the chi-square difference be significant it would imply that the factor loadings differ across samples [i.e. the item calibration is different across samples]. The manner in which the items are calibrated to respond to an increase in the latent variable they were designed to represent would thus be different across the two samples. This would constitute a somewhat disappointing outcome as it would imply that the PI fails in its attempt to measure the underlying unit performance construct in the same way across samples from the same population. A Chi-square difference test is used to determine the difference between the Satorra-Bentler Chi-square values for the model with the factor loading equality constraints imposed and for the model with without equality constraints, taking into account the accompanying loss of degrees of freedom (Mels, 2003). If the difference in Chi-square values is significant ($p<0,05$) the null hypothesis should be rejected, and the model may be considered to differ across the two samples in the manner in which the item parcels load on the latent variables. A difference in Chi-square values that is not significant, on the other hand, would indicate that the null hypotheses that the lambda-X matrices are the same across the two samples couldn't be rejected (Mels, 2003). The results of the Chi-square difference test are given in Table 29.

TABLE 29

CHI-SQUARE DIFFERENCE FOR TEST OF METRIC INVARIANCE

Hypothesis	Chi-square value	Degrees of freedom	Critical χ^2 (p=0,05)
Model with factor loadings set to be equal (H_0)	276,36	168	
Model with no parameters constrained (H_a)	256,77	152	
Difference	19,59	16	26,30

(*= significant at the p<0,01 level; ** = significant at p< 0,05)

The difference between the Chi-square values is not significant which indicates the fit does not deteriorate significantly under the factor loading equality constraints and that the null hypothesis of equal lambda_X matrices can not be rejected. This suggests that the extent to which the content of each item is being perceived and interpreted in exactly the same way across samples (Byrne & Watkins, 2003). The PI measurement model thus displays metric invariance. This is a satisfying result as, without at least partial metric invariance, the envisaged subsequent tests of structural invariance would have been questionable. Further tests of measurement invariance are,

however, now required to determine which model parameter estimates show significant variance when the measurement model is applied across both samples.

4.7.3 Test for equivalence of factor covariances

Having established that metric invariance justifies the use of further tests of measurement invariance to determine if the difference in parameter estimates that exist in the model are due to differences in factor variances and covariances and/or error variances. Testing for the equivalence of factor covariances between groups tests the null hypothesis that the phi matrices are invariant across both samples ($\Phi^g_{ij} = \Phi^g_{ij}$). The main diagonal in the phi matrix represents the latent variable variance terms. "Factor variances represent the dispersion of the latent variables (ξ_j) and thus represent variability of the construct continua within the samples" (Vandenberg & Lance, 2000, p. 39). The off-diagonal elements in the phi matrix represent the covariance between the latent variables.

Using the Chi-square difference test, the model fit with factor variances set equal to one and covariances (i.e., the off-diagonal elements of Φ) constrained to be equal across groups is compared to the unconstrained model. Failure to reject the null hypothesis that $\Phi^g_{ij} = \Phi^g_{ij}$ would imply that both samples use "equivalent ranges of the construct continuum to respond to the indicators reflecting the construct" (Vandenberg & Lance, 2000, p.39) and that the latent variables covary in the same manner across samples. On the other hand, rejection of the null hypothesis of full phi matrix equivalence would signal differences in the covariance terms (Φ_{ij}). The results of the Chi-square difference test for the equivalence of the full phi matrix across the two samples are given in Table 30.

TABLE 30

CHI-SQUARE DIFFERENCE TEST - EQUIVALENCE OF FACTOR VARIANCES AND COVARIANCES

Hypothesis	Chi-square value	Degrees of freedom	Critical χ^2 (p=0,05)
Model with factor variances and covariances set to be equal (H_0)	303,18	180	
Model with no parameters constrained (H_a)	256,77	152	
Difference	46,41**	28	41,34

(*= significant at the p< 0,01 level; ** = significant at p< 0,05)

The difference between the Chi-square values is significant (p<0,05). This indicates that the factor covariances and/or variances may not be considered invariant across both samples and that, to some extent, both samples do not respond to the indicator variables in exactly the same way. Nonetheless, the difference between the Chi-square values is insignificant at p<0,01. This implies that variance between the samples factor covariances and/or covariances may indeed be quite small. It thus remains to be seen to what extent the variance in the measurement model that was established through previous tests may be attributed to the non-equivalence in the phi matrices across samples only, or may also be due to non-equivalence in error variances. This question may be answered by testing for the equivalence of error variance across samples.

4.7.4 Test for equivalence of error variances

To shed more light on the question of where variance exists in the measurement model, a test is required to test the null hypothesis of equal variance in the error terms associated with the indicator variables across groups ($\theta_{\delta}{}^{g}{}_{j} = \theta_{\delta}{}^{g}{}_{j}{}'$). This test involves conducting a Chi-square difference test that compares the model fit when the error variances of the like factor pairs are constrained to be equal across groups, to the fit of the model with no parameters constrained to be equal.

Failure to reject the null hypothesis would provide evidence that both samples respond to the indicator variables in an equivalent manner, in that the no significant variance exists across samples in terms of the error terms associated with the indicator variables. Rejection of the null hypothesis would, however, imply some of the variance in the measurement model fit between the two samples, may be attributed to non-equivalent error variance across samples.

The results of the Chi-square difference test are given in Table 31. The difference between the Chi-square values is significant at p<0,05. This indicates that the error variances may not be considered invariant across both samples. Nonetheless, as the Chi-square difference is insignificant at p<0,01 that variance between the error variances across samples may indeed be quite small.

In sum, when considering the information presented above collectively, the variance in the measurement model fit across Sample A and Sample B seems to exist in both the measurement error variances and factor covariances matrices. Furthermore, this variance appears to be fairly

small for both the error variances and factor covariances due to the insignificance of the difference in Chi-square values at p<0,01.

TABLE 31

CHI-SQUARE DIFFERENCE TEST – EQUIVALENCE OF ERROR VARIANCES

Hypothesis	Chi-square value	Degrees of freedom	Critical χ^2 (p=0,05)
Model with error variances set to be equal (H_0)	287,87	168	
Model with no parameters constrained (H_a)	256,77	152	
Difference	31,10**	16	26,30

(*= significant at the p< 0,01 level; ** = significant at p< 0,05)

CHAPTER 5
DISCUSSION, CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE RESEARCH

The objective of this cross-validation study was to establish the extent to which the PI measurement model may be considered measurement invariant across two independent samples from the target population. A series of measurement invariance tests were used to test the stability of the model parameter estimates in order to determine the source of the variance and to what extent the measurement model may be considered measurement invariant or not at all (Vandenberg & Lance, 2000). The results of the study failed to find support for the hypothesis that the measurement model fits the data of both samples in exactly the same way and thus the PI could not be said to display full measurement invariance. As discussed previously, this finding is expected given the nature of social science research (Steenkamp & Baumgartner, 1998) and the omnibus test was conducted in the attempt to follow a prudent research process. Moreover, the problem that the two samples were not two probability samples from the same target population but rather two convenience non-probability samples should also be borne in mind.

The measurement model did cross-validate successfully under the configural invariance condition, which requires the *apriori* pattern of free and fixed factor loadings imposed on the measure's components to be constrained. This indicates that some of the measurement model parameter estimates show significant variance when the measurement model is applied across both samples. Subsequent tests of measurement invariance were used to determine the source of non-equivalence evident in the omnibus measurement invariance test. In contrast to the above results, the factor loadings of item parcels on latent variables was found to be equivalent across both samples, supporting the conclusion that the measurement model displays at least metric invariance across the samples. Finding at least metric invariance is a satisfying outcome as it indicates the content of each item is perceived and interpreted in a similar manner across samples from the target population.

Further investigation revealed that the differences in parameter estimates across the samples exist in both the error variances and the factor covariances ($p<0,05$) although these differences may be argued to be negligible as the differences in the parameter estimates for both the error variances and the factor covariances becomes non-significant at the $p<0,01$ level. What is also important

to note is that the variance due to the error variances and the variance due to factor covariances appear equal. These results suggest that the non-equivalence that exists in both the error variances and factor covariances is not very large, especially as the $p<0,05$ critical value is not surpassed by much in both cases. When considered in combination, these results may be viewed as quite satisfactory as they indicate that the measurement model does not appear to vary greatly when fitted to data from the two samples.

As part of ongoing research of the leadership-for-performance range of measures designed by Spangenberg and Theron (2004), this study takes the initial step towards establishing the degree of confidence with which the PI may be used across different groups within the target population. As such, these results, in particular the result of metric invariance, provides sufficient support for the conclusion that the PI is successful in its attempt to measure the underlying unit performance construct in almost exactly the same way across these samples. Although the results also indicate that some differences exist when the measurement model is fitted to both samples simultaneously and should be taken into account when examining the structural invariance of the PI, these results do not appear to threaten the conclusion that the PI is a credible measure of the work unit performance construct it was intended to measure.

These results, therefore, justify continued research that seeks evidence of structural model fit through tests of structural invariance, or other research that examines the structural relationships between the latent dimensions of the PI and the possible refinement of the PI model as suggested by the measurement model modification indices. Of particular interest are the previous research findings of Henning *et al.* (2003) and Theron *et al.* (2004) that suggest Adaptability, Climate and Capacity do not have a significant impact on Production & Efficiency. As this study has established at least metric invariance of the PI, it therefore provides some basis of confidence for findings of subsequent research that links the leadership behaviour to work unit performance as measured by the PI.

Lastly, it is important to acknowledge certain limitations of this study. Firstly, invariance across the samples used in this study may not be assumed to mean invariance across qualitatively different groups within the target population or across samples from other populations. This study should, therefore, be replicated across other samples from the target population in order to further establish the measurement invariance of the PI.

Secondly, to convincingly demonstrate that the PI functions effectively within the target population this study should have employed the data from two independent random samples taken from the same population rather than non-probability samples. A consideration for future studies is to ensure the samples better represent the target population through inclusion of multiple industries and companies. As the samples in this study cannot be said to constitute representative sections of the population of work units, it is not possible to reach a definitive conclusion that the PI can be used for all organisations and industries across the target population using these results. It is recommended that the study be repeated on the same data but that the two samples be combined and randomly divided into two samples. These two samples would still not constitute probability samples from the target population but at least the current fear that the lack of full measurement invariance had been due to systematic differences between the two non-probability samples would have been allayed.

A third potential limitation of this study that needs to be discussed relates to how factor loadings were fixed. The measurement model relating the eight latent unit performance dimensions measured by the PI to the sixteen item parcels can be expressed in terms of the following regression equation:

$$X = \tau_X + \Lambda_X\xi + \delta \qquad 2$$

Where:

X represents a 16x1column vector X of item parcels;

τ_X a 16x1 column vector of regression intercept terms;

Λ_X a 16x8 factor loading matrix of regression slopes;

ξ a 8x1 column vector of exogenous latent variables; and

δ a 16x1 column vector of measurement error terms.

In equation 2 E[δ] and $\rho[\xi,\delta]$ are assumed to be zero but it is not assumed that E[ξ] or E[X] are zero. In the present study the latent variables were assumed to have a mean of zero. The observed item parcels, moreover were assumed to be deviations around the mean. The measurement model depicted as equation 2 above thereby reduces to equation 1 depicted on p. 51. Since τ_X represents the means indicator variable scores when the latent variable it expresses is zero, the vector of intercept terms reduces to a vector of zeros and thus could be omitted from equation 1. The present study thus did not examine the possibility of differences in intercepts of

the regression of item parcels on the latent variables they were meant to represent. Equation 2 would, however, suggest that H_0: $\tau^{g}_{X}=\tau^{g'}_{X}$ is a meaningful and relevant hypothesis to examine measurement equivalence across independent random samples from the same population (Chan, 2000; Vandenberg & Lance, 2000). The critical question is whether the manner in which observed item responses relate to the underlying latent variable remains the same across the two samples. Only if the regression of the item parcel on the latent variable fully coincides in terms of intercept and slope would the same latent variable/trait inference from the same observed score obtained by individuals from two samples be justified.

Moreover, in fitting the measurement model to the two separate samples the present study only freed the off-diagonal elements of the phi matrix to be estimated but not the diagonal elements of the phi matrix. The latent variables were by implication assumed to be standardized (0,1) variables so that the latent variable variances were thereby fixed to one. No unstandardiszed phi matrix with variance-covariance estimates is thus obtained but rather only a standardized phi matrix with estimates of the correlations between latent variables. The latent variables variances are by definition then equal to unity. The scale on which all latent variables are expressed is thereby calibrated in terms of standard deviation units. When fitting a measurement model to a single group this solution to the problem that latent variables have no inherent scale seems preferable to fixing the factor loading of on indictor variable to unity for each latent variable.

When examining possible differences in measurement model parameter estimates across groups, however, literature on the use of structural equation modelling in the evaluation of measurement equivalence (De Bruin, personal communication, 9 November 2006; Mels, 2003; Mels, personal communication, 28 March 2006; Nesselroade & Thompson, 1995) seems to favour the procedure of fixing the loading of the first indicator variable of each latent variable to one. The scale of the first indicator variable thereby sets the scale for the latent variable. To follow the standard procedure of fixing the variance of the latent variables to unity would from the outset assume that latent variable variances are equal to one and equal across groups (De Bruin, personal communication, 9 November 2006; Mels, personal communication, 28 March 2006). In a cross-validation study with two independent random samples from the same target population this could conceivably be the case. Although the assumption might under these circumstances be quite reasonable it nonetheless to a certain degree negates the spirit and objective of a cross validation study to empirically examine the invariance of all measurement model parameter estimates across samples. When however the two samples in question constitute independent

but not random samples from the same target population the decision to fix the latent variable variances to unity becomes somewhat more contentious (De Bruin, personal communication, 9 November 2006).

The present study used two item parcels to operationalise each of the latent PI dimensions. Fixing the factor loading of the first indicator variable of each latent variable to one for both groups to thereby scale the latent variables to the scale of the first indicator variable would, therefore have resulted in half the factor loadings in Λ_X being fixed to one and only half of the factor loading matrix being evaluated in terms of metric invariance. This moreover raises the concern that those factor loadings being fixed might in fact be concealing difference in regression slopes across samples.

The decision on whether to fix the factor loading of the first indicator variable of each latent variable to one or to fix the variance of the latent variables to unity is therefore not that clear-cut. Moreover no single study should probably aspire to provide a final, definitive answer to the question whether measurement invariance exists across independent random samples from the same target population. Nonetheless it probably would have been more prudent if the present study would have chosen to fix the factor loading of the first indicator variable of each latent variable to one rather than to fix the variance of the latent variables to unity and two set the latent variables free in the fully unconstrained solution. It is thus recommended that, in repeating the study on the same data as suggested above, the factor loading of the first indicator variable of each latent variable should be fixed to one rather than fixing the variance of the latent variables to unity and that the main diagonal of the full phi matrix should be estimated [i.e., also the main diagonal]. Therefore when establishing whether the measurement model, when fitted to the two samples simultaneously in a multi-group analysis with no freed parameters constrained, display reasonable fit (Step 2; paragraph 2.2), the latent variable variances will form part of the freed model parameters. Likewise when establishing whether the measurement model demonstrated acceptable fit when fitted to the two samples simultaneously in a multi-group analysis with all freed parameters constrained to be equal across the samples (Step 2; paragraph 2), the latent variable variances will again form part of the freed model parameters. This will remain the case across all the steps outlined earlier.

In the re-analysis of a random split of the current combined sample the use of more sophisticated approaches to the dimensionality analysis of the subscales should also be explored in addition to

the use of principle axis factor analysis and the extraction of factors with eigenvalues greater than unity. Specifically exploratory principal factor analysis could be used to fit one and two factor models and to inspect the residuals, RMSEA and ECVI for the two models.

In conclusion, and despite the shortcomings outlined above, this measurement invariance study provides reasonable, albeit limited and tentative, evidence that the PI measurement model demonstrates partial measurement invariance across these two samples from the same population.

REFERENCES

Anastasi, A. & Urbina, A. (1997). *Psychological testing.* USA, NJ: Prentice-Hall Inc.

Babbie, E. & Mouton, J. (2001). *The practice of social research.* Oxford University Press: Oxford.

Bandalos, D. (2002). The effects of item parceling on goodness-of-fit and parameter estimate bias in structural equation modeling. *Structural Equation Modeling, 9*(1), 78-102.

Bandalos, D.L., Geske, J.A. & Finney, S.J. (2001). A model of statistics performance based on achievement goal theory. *Journal of Educational Psychology, 95*(3), 604-616.

Bass, B.M, Avolio, B., Jung, D. & Berson, Y. (2003). Predicting unit performance by assessing transformational and transactional Leadership. *Journal of Applied Psychology, 88*(2), 207–218.

Bentler, P.M. & Chou, C. (1987). Practical issues in structural modeling. *Sociological Methods & Research, 16*(1), 78-118.

Bernstein, I.H., Teng, G., Grannemann, B.D. & Garbin, C.P. (1987). Invariance in the MMPI's component structure. *Journal of Personality Assessment, 51*(4), 522-532.

Bollen, K.A. & Long, J.S. (1993). *Testing Structural Equation Models.* USA: SAGE Publications, Inc.

Browne, M.W. & Cudeck, R. (1993). Alternative ways of assessing model fit. In Bollen, K.A. & Long, J.S. (Eds.). *Testing structural equation models.* Newbury Park: Sage Publications.

Bunderson, J. S. & Sutcliffe, K.L. (2002). Comparing alternative conceptualisations of functional diversity in management teams: process and performance effects. *Academy of Management Journal, 45*(5), 875-893.

Bunderson, J. S. & Sutcliffe, K.L. (2003). Management team learning orientation and business unit performance. *Journal of Applied Psychology, 88*(3), 552–560

Byrne, B.M. (1998). *Structural equation modelling with LISREL, PRELIS, and SIMPLIS: basic concepts, applications and programming.* Mahwah, NJ: Lawrence Erlbaum Associates.

Byrne, B.M.. & Watkins, D. (2003). The issue of measurement invariance revisited. *Journal of Cross-Cultural Psychology, 34*(2), 155-175.

Chan, D. (2000). Detection of differential item functioning on the Kirton Adapttion-Innovation Inventory using multiple-group mean and covariance structure analysis. *Multivariate Behavioral Research, 35*(2), 169-199.

Cheung, G.W. & Rensvold, R.B. (2000). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management, 25,* 1-27.

Conger, J.A. & Kanungo, R.N. (1998). *Charismatic leadership in organisations.* Thousand Oaks, CA: Sage Publications.

Davey, A., Shanahan, M.J. & Schafer, J.L. (2001). Correcting for selective nonresponse in the national longitudinal survey of youth using multiple imputation. *Journal of Human Resources, 36*(3), 500-519.

Diamantopoulos, A. & Siguaw, J.A. (2000). *Introducing LISREL*. London: Sage Publications.

Du Toit, M. & Du Toit, S. (2001). *Interactive LISREL: User's guide*. Lincolnwood, IL: Scientific Software International.

Du Toit, S.H.C. & Mels, G. (2002). Supplementary notes on multiple imputation. http://www.ssicentral.com/other/impute.htm.

Durvasula, S., Andrews, J.C., Lysonski, S., & Netemeyer, R.G. (1993). Assessing the cross-national applicability of consumer behaviour models: a model of attitude toward adverting in general. *Journal of Consumer Research, 19*, 626-636.

Enders, C.K. & Bandalos, D.L. (2001). The relative performance of full maximum likelihood estimation for missing data in structural equation modelling. *Structural Equation Modeling, 8*(3), 430-457.

Fay, D., Luhrmann & Kohl, C. (2004). Proactive climate in a post-reorganization setting: When staff compensate managers' weakness. *European Journal of Work and Organisational Psychology, 13*(2), 241-267

Forbes, D.P. (1998). Measuring the unmeasurable: empirical studies of non-profit organisational effectiveness from 1977-1997. *Nonprofit and Voluntary Sector Quarterly, 27*, 183-202.

Gelade, G. & Ivery, M. (2003). The impact of human resource management and work climate on organisational performance. *Personnel Psychology, 56,* 383-404.

Gibson, C.B. & Birkinshaw J. (2004). The antecedents, consequences, and mediating role of organisational ambidexterity. *Academy of Management Journal, 47*(2), 209–226.

Gibson, J.L., Ivancevich, J.M. & Donelly, J.H. (1991). *Organisations*. Boston: Irwin Publishers.

Globerson, S. & Riggs, J.L. (1989). Multi-performance measures for better operational control. *International Journal of Productivity, 27*(1), 187-194.

Green, J.C., Madjidi, F., Dudley, T.J. & Gehlen, F.L. (2001). Local unit performance in a national nonprofit organisation. *Nonprofit Management and Leadership, 11*(4), 459-476.

Hagvet, K.A. & Nasser, F.M. (2004). How well do item parcels represent conceptually defined latent constructs? A two-facet approach. *Structural Equation Modeling, 11*(2), 168-193.

Hagvet, K.A. & Zuo, L. (2000). Conceptual and empirical components of an internal domain study: an illustration in terms of the Achievement Motives Scale. *Scandinavian Journal of Educational Research, 44*(1), 49-78.

Hall, R.J., Snell, A.F. & Foust, M. (1999). Item parceling strategies in SEM: Investigating the subtle effects of unmodeled secondary constructs. *Organizational Research Methods, 2*(3), 233-251.

Henning, R., Theron, C.C. & Spangenberg, H.H. (2003). An investigation into the internal structure of the unit performance construct as measured by the Performance Index (PI). *Manuscript submitted to the South African Journal of Industrial Psychology.*

Holt, J.K. (2003). *Item parcelling in structural equation models for optimum solutions.* Paper presented at the 2004 Annual Meeting of the Mid-Western educational Research Association.

House, R.J. (1998). Leadership research: some forgotten, ignored, or overlooked findings. In J.G. Hunt, B.R. Boliga, H.P. Dachler and C.A.Schriesheim (Eds.): *Emerging leadership vistas.* Lexington MA: Lexington Books.

Hu, L.T. & Bentler, P.M. (1995°. Evaluating model fit. In Hoyle, R.C. (Ed.): *Structural equation modelling: concepts, issues and applications.* Thousand Oaks: Sage Publications.

Hulin, C.L., Drasgow F. & Parsons C.K. (1983). *Item response theory: application to psychological measurement.* Homewood, Ill.: Jones-Irwin Publishers.

Information Technology Services (2005). General FAQ #25: Handling missing or incomplete data. http://www.utexas.edu/its/rc/answers/general/gen25.html. [Retrieved on 30/08/2005].

Javidan, M. & Waldman, D.A. (2003). Exploring charismatic leadership in the Public sector: Measurement and consequences. *Public Administration Review, 63*(2), 229-242.

Jöreskog, K.G. & Sörbom, D. (1993). *LISREL 8: Structural equation modeling with SIMPLIS command language.* Chicago: Scientific Software International.

Jöreskog, K.G. & Sörbom, D. (1996a). *PRELIS 2: User's reference guide.* Chicago: Scientific Software International.

Jöreskog, K.G. & Sörbom, D. (1996b). *LISREL 8: User's reference guide.* Chicago: Scientific Software International.

Jöreskog, K.G. & Sörbom, D. (1998). *Structural equation modelling with the SIMPLIS command language.* Chicago: Scientific Software International.

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

Kaplan R. & Norton, D. (1992). Using the Balanced Scorecard as a strategic management system. *Harvard Business Review, 74*(1), 71-19.

Kaplan, D. (1995). Statistical power in structural equation modeling. In R.H. Hoyle (Ed.): *Structural equation modeling: Concepts, issues and application.* Thousand Oaks, CA: Sage Publications.

Kaplan, D. (2000). Structural equation modeling: foundations and extensions. Thousand Oaks, CA: Sage Publications.

Kaplan, R.S. & Norton, D.P. (1996). The Balanced Scorecard: Translating strategy into action. Boston, Mass.: Harvard Business School Press.

Kelloway, E.K. (1998). *Using LISREL for structural equation modelling: a researcher's guide.* Thousand Oaks, CA: SAGE Publications, Inc.

Kerlinger, F. & Lee, H. (2000). *Foundations of behavioural research.* Fort Worth, Tex.: Harcourt College Publishers.

Kolb, D.A. (1996). A comparison of leadership behaviours and competencies in high- and average- performance teams. *Communication Reports, 9*(2), 173-185.

Lipe, M.G. & Salterio, S.E. (2000). Balanced Scorecard: Judgemental effects. *The Accounting Review, July,* 283-293.

Little, T.D., Cunningham, W.A., Shahar, G., & Widaman, K.F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling, 9*, 151-173.

Loughry, M.L. (2002). Coworkers are watching: Performance implications of peer monitoring. *Academy of Management Proceedings,* 1-6.

Lubke, G.H. & Muthen, B.O. (2004). Applying multi-group confirmatory factor models for continuous outcomes to Likert scale data complicates meaningful group comparisons. *Structural Equation Modeling, 11*(4), 514-534.

MacCallum, R.C. 1995. Model specification: procedures, strategies and related issues. In Hoyle, R.H. (Ed.). *Structural equation modelling: concepts, issues and applications.* Thousand Oaks, CA: Sage Publications.

Marsh, H.W., Hau, K.T., Balla, J.R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research, 33*, 181-220.

Mavondo, F., Gabbott, M. & Tsarenko, Y. (2003). Measurement invariance of marketing instruments: An implication across countries. *Journal of Marketing Management, 19*, 523-540.

Mels, G. (2003). *A workshop on structural equation modeling with LISREL 8.54 for Windows.* Chicago: Scientific Software International.

Nasser, F. & Takahashi, T. (2003). The effect of using item parcels on ad hoc goodness-of-fit indices in confirmatory factor analysis: An example using Sarason's Reactions to Tests. *Applied Measurement in Education, 16,* 75-97.

Nesselroade, J.R. & Thompson, W.W. (1999). Selection and related threats to group comparisons: an example comparing factor structures of higher and lower ability groups of adult twins. *Psychological Bulletin, 117*, 271-284.

Nicholson, N. & Brenner, S.O. (1994). Dimensions of perceived organisational performance: tests of a model. *Applied Psychology: an International Review, 43*(1), 69-108.

Peterson, M.F., *et al. (*1995). Role conflict, ambiguity, and overload: a 21-nation study. *Academy of Management Journal, 38*(2), 429-452.

Pigott, T.D. (2001). A review of methods for missing data. *Educational Research & Evaluation, 7*(4), 353-383.

Pousette, A. & Hanse, J.J. (2002). Job characteristics as predictors of ill-health and sickness absenteeism in different occupational types: A multi-group structural equation modelling approach. *Work and Stress, 16*(3), 229-250.

Raghunathan, T.E. (2004). What do we do with missing data? Some options for analysis of incomplete data. *Annual Review of Public Health, 25*(1), 99-117.

Roth, P.L. (1994). Missing data: A conceptual review for applied psychologists. *Personnel Psychology, 47*(3), 537-560.

Safferstone, M.J. (1999). Did you hear the one about…? Leading with humor pays dividends. *Academy of Management Executive, November,* 103-104.

Sale, M.L. & Inman, R.A. (2003). Survey-based comparison of performance and change in performance of firms using traditional manufacturing, JIT and TOC. *International Journal of Production Research, 41*(4), 829-844.

Sartori, N. Salvan, A., & Thomaseth, K. (2005). Multiple imputation of missing values in a cancer mortality analysis with estimated exposure dose. *Computational statistics & Data Analysis, 49*(3), 937-953.

Schafer, J.L. & Olsen, M.K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioural Research, 33*(4), 545-571.

Schafer, J.L. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research, 8*, 3-15.

Schumacker, R.E. & Lomax, R.G. (1996). *A beginner's guide to structural equation modeling.* Mahaw, NJ: Lawrence Erlbaum Associates, Publishers.

Spangenberg, H.H. & Theron, C.C. (1997). Developing a Performance Management Audit Questionnaire. *South African Journal of Psychology, 27*(3), 143-150.

Spangenberg, H.H. & Theron, C.C. (2002). Development of a uniquely South African leadership questionnaire. *South African Journal of Psychology*, *32*(2), 9-25.

Spangenberg, H.H. & Theron, C.C. (2004). Development of a performance measurement questionnaire for assessing organisational work unit effectiveness. *South African Journal of Industrial Psychology, 30*(1), 19-28.

SPSS 13 for Windows. (2005). SPSS Inc. http://www.spss.com/

Steenkamp, J. & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research, 25*, 78-90.

Stewart, D. (2001). Factor analysis. *Journal of Consumer Psychology, 10*(1&2), 75-82.

Switzer, F.S., Roth, P.L., & Switzer, D.M. (1998). Systematic data loss in HRM settings: A Monte Carlo analysis. *Journal of Management, 24*(6), 763-779.

Tabachnick, B.G. & Fidell, L.S. (1989). *Using multivariate statistics*. New York: Harper & Row.

Theron, C., Spangenberg, H. & Henning, R. (2004). An elaboration of the internal structure of the unit performance construct as measured by the Performance Index (PI). *Management Dynamics, 13*(2), 35-52.

Vandenberg, R.J. & Lance, C.E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices and recommendations for organizational research. *Organizational Research Methods, 2*, 4-69.

Vandenberg, R.J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. *Organizational Research Methods, 5*(2), 139-158.

Watson, A., & Wooldridge, B. (2005). Business unit manager influence on corporate-level strategy formulation. *Journal of Managerial Issues, 18*(2), 147-161.

Yukl, G. (2002). *Leadership in Organizations*. Prentice Hall International: New Jersey.

APPENDIX 1

TABLE 27: DESCRIPTIVE STATISTICS FOR SAMPLE A

		Item1	Item2	Item3	Item4	Item5	Item6	Item7	Item8	Item9	Item10	Item11	Item12	Item13	Item14	Item15	Item16	Item17	Item18	Item19
N	Valid	277	277	277	277	277	277	277	277	277	277	277	277	277	277	277	277	277	277	277
Mean		3,755	3,704	4,051	3,440	3,870	3,542	3,487	3,473	3,415	3,552	3,646	3,596	3,563	3,079	3,733	3,513	3,534	3,704	3,610
Std. Deviation		0,736	0,807	0,745	0,860	0,900	0,749	0,980	0,998	0,991	0,922	0,923	1,054	1,022	1,060	1,008	0,962	0,915	0,916	0,985
Variance		0,541	0,651	0,555	0,740	0,809	0,561	0,961	0,997	0,983	0,850	0,853	1,111	1,044	1,124	1,015	0,925	0,837	0,840	0,971
Skewness		*-0,678**	-0,119	*-0,505**	-0,020	-0,282	-0,012	-0,127	-0,266	-0,145	-0,225	*-0,408**	*-0,609**	*-0,439**	-0,104	*-0,300**	-0,073	-0,174	*-0,489**	*-0,482**
Std. Error of Skewness		0,146	0,146	0,146	0,146	0,146	0,146	0,146	0,146	0,146	0,146	0,146	0,146	0,146	0,146	0,146	0,146	0,146	0,146	0,146
Kurtosis		*1,704**	0,137	0,367	0,483	-0,409	-0,031	-0,236	-0,098	-0,279	-0,045	0,331	0,143	-0,040	-0,328	-0,473	-0,318	-0,020	0,361	0,199
Std. Error of Kurtosis		0,292	0,292	0,292	0,292	0,292	0,292	0,292	0,292	0,292	0,292	0,292	0,292	0,292	0,292	0,292	0,292	0,292	0,292	0,292

		Item20	Item21	Item22	Item23	Item24	Item25	Item26	Item27	Item28	Item29	Item30	Item31	Item32	Item33	Item34	Item35	Item36	Item37	Item38
N	Valid	277	277	277	277	277	277	277	277	277	277	277	277	277	277	277	277	277	277	277
Mean		3,513	3,657	3,379	3,585	2,805	2,809	3,379	3,874	3,816	3,690	3,675	3,484	3,563	3,386	3,112	3,581	3,549	3,632	3,018
Std. Deviation		0,923	0,941	0,806	0,858	0,977	1,044	0,988	0,941	0,981	0,962	0,870	0,923	0,929	0,892	0,908	1,028	0,953	0,937	0,961
Variance		0,852	0,886	0,649	0,736	0,955	1,090	0,975	0,886	0,962	0,925	0,756	0,852	0,863	0,796	0,824	1,056	0,908	0,878	0,924
Skewness		-0,246	*-0,396**	0,168	-0,162	-0,023	-0,052	-0,251	*-0,559**	*-0,505**	*-0,352**	*-0,416**	-0,161	*-0,500**	-0,131	-0,018	*-0,301**	-0,178	-0,214	0,112
Std. Error of Skewness		0,146	0,146	0,146	0,146	0,146	0,146	0,146	0,146	0,146	0,146	0,146	0,146	0,146	0,146	0,146	0,146	0,146	0,146	0,146
Kurtosis		0,040	-0,140	0,219	0,067	-0,038	-0,403	-0,076	0,159	-0,136	-0,172	0,496	-0,230	0,232	0,133	0,168	-0,420	-0,351	-0,468	0,147
Std. Error of Kurtosis		0,292	0,292	0,292	0,292	0,292	0,292	0,292	0,292	0,292	0,292	0,292	0,292	0,292	0,292	0,292	0,292	0,292	0,292	0,292

		Item39	Item40	Item41	Item42	Item43	Item44	Item45	Item46	Item47	Item48	Item49	Item50	Item51	Item52	Item53	Item54	Item55	Item56
N	Valid	277	277	277	277	277	277	277	277	277	277	277	277	277	277	277	277	277	277
Mean		3,097	3,296	3,343	3,292	3,422	3,357	3,386	3,538	3,361	3,765	3,560	3,505	3,783	3,440	3,220	3,509	3,610	3,227
Std. Deviation		0,997	0,916	0,949	0,954	0,888	1,129	0,838	0,942	1,090	0,884	1,036	0,837	0,895	0,971	0,736	0,899	0,872	1,054
Variance		0,994	0,840	0,900	0,911	0,788	1,274	0,702	0,887	1,188	0,782	1,073	0,700	0,801	0,943	0,542	0,809	0,760	1,111
Skewness		-0,020	*-0,365**	-0,247	0,015	-0,186	-0,266	0,066	-0,110	*-0,451**	-0,283	*-0,515**	-0,241	-0,263	-0,022	-0,208	-0,193	-0,276	-0,167
Std. Error of Skewness		0,146	0,146	0,146	0,146	0,146	0,146	0,146	0,146	0,146	0,146	0,146	0,146	0,146	0,146	0,146	0,146	0,146	0,146
Kurtosis		-0,157	0,220	0,044	-0,211	0,193	-0,435	0,329	-0,196	-0,204	-0,193	0,034	0,352	-0,431	-0,191	*1,291**	-0,069	0,348	-0,124
Std. Error of Kurtosis		0,292	0,292	0,292	0,292	0,292	0,292	0,292	0,292	0,292	0,292	0,292	0,292	0,292	0,292	0,292	0,292	0,292	0,292

* significant (p<0,05)

APPENDIX 2

TABLE 28: DESCRIPTIVE STATISTICS FOR SAMPLE B

		Item1	Item2	Item3	Item4	Item5	Item6	Item7	Item8	Item9	Item10	Item11	Item12	Item13	Item14	Item15	Item16	Item17	Item18	Item19
N	Valid	375	375	375	375	375	375	375	375	375	375	375	375	375	375	375	375	375	375	375
Mean		3,776	3,739	4,035	3,563	3,867	3,608	3,605	3,488	3,507	3,576	3,536	3,643	3,555	3,357	3,616	3,552	3,520	3,605	3,488
Std. Deviation		0,741	0,778	0,729	0,753	0,861	0,745	1,015	0,880	0,877	0,895	0,833	0,884	0,911	0,995	0,926	0,914	0,833	0,871	0,898
Variance		0,549	0,605	0,531	0,568	0,742	0,554	1,031	0,775	0,769	0,801	0,693	0,781	0,831	0,990	0,857	0,836	0,694	0,758	0,807
Skewness		*-0,251**	0,184	*-0,387**	*0,389**	-0,144	0,231	*-0,258**	-0,105	0,135	-0,140	-0,185	-0,007	*-0,429**	-0,142	0,063	-0,155	0,118	-0,119	-0,142
Std. Error of Skewness		0,126	0,126	0,126	0,126	0,126	0,126	0,126	0,126	0,126	0,126	0,126	0,126	0,126	0,126	0,126	0,126	0,126	0,126	0,126
Kurtosis		*0,508**	*-0,806**	-0,120	-0,267	*-0,672**	-0,262	*-0,540**	0,055	*-0,585**	-0,109	0,290	*-0,674**	0,242	-0,265	*-0,656**	-0,221	-0,296	-0,426	0,038
Std. Error of Kurtosis		0,251	0,251	0,251	0,251	0,251	0,251	0,251	0,251	0,251	0,251	0,251	0,251	0,251	0,251	0,251	0,251	0,251	0,251	0,251

		Item20	Item21	Item22	Item23	Item24	Item25	Item26	Item27	Item28	Item29	Item30	Item31	Item32	Item33	Item34	Item35	Item36	Item37	Item38
N	Valid	375	375	375	375	375	375	375	375	375	375	375	375	375	375	375	375	375	375	375
Mean		3,512	3,744	3,299	3,403	2,680	2,899	3,269	3,755	3,712	3,568	3,637	3,477	3,475	3,317	3,187	3,408	3,347	3,581	3,280
Std. Deviation		0,859	0,880	0,844	0,847	0,922	0,959	0,962	0,877	0,929	0,865	0,851	0,846	0,787	0,826	0,941	0,917	0,854	0,842	0,800
Variance		0,737	0,774	0,713	0,717	0,849	0,920	0,925	0,769	0,863	0,749	0,724	0,715	0,619	0,682	0,885	0,841	0,730	0,709	0,641
Skewness		-0,076	-0,140	0,086	-0,024	0,080	0,131	0,001	-0,243	-0,221	-0,087	-0,121	0,085	0,167	0,010	0,046	0,064	-0,032	-0,082	0,018
Std. Error of Skewness		0,126	0,126	0,126	0,126	0,126	0,126	0,126	0,126	0,126	0,126	0,126	0,126	0,126	0,126	0,126	0,126	0,126	0,126	0,126
Kurtosis		-0,269	-0,437	0,118	-0,007	-0,015	-0,226	-0,379	-0,314	*-0,551**	-0,177	-0,348	-0,331	-0,226	0,058	-0,360	-0,450	0,129	-0,055	*0,670**
Std. Error of Kurtosis		0,251	0,251	0,251	0,251	0,251	0,251	0,251	0,251	0,251	0,251	0,251	0,251	0,251	0,251	0,251	0,251	0,251	0,251	0,251

		Item39	Item40	Item41	Item42	Item43	Item44	Item45	Item46	Item47	Item48	Item49	Item50	Item51	Item52	Item53	Item54	Item55	Item56
N	Valid	375	375	375	375	375	375	375	375	375	375	375	375	375	375	375	375	375	375
Mean		3,192	3,136	3,256	3,395	3,493	3,512	3,368	3,387	3,216	3,576	3,264	3,616	3,683	3,421	3,365	3,509	3,261	3,259
Std. Deviation		0,869	0,995	0,886	0,823	0,804	0,964	0,796	0,903	0,964	0,800	1,068	0,758	0,852	0,780	0,684	0,807	0,834	1,021
Variance		0,754	0,989	0,785	0,678	0,646	0,930	0,634	0,815	0,929	0,641	1,141	0,574	0,725	0,608	0,468	0,652	0,696	1,043
Skewness		0,208	-0,194	0,195	0,023	*0,270**	-0,007	-0,216	-0,009	-0,246	0,096	*-0,344**	0,214	0,136	*0,298**	*0,596**	0,200	0,144	-0,051
Std. Error of Skewness		0,126	0,126	0,126	0,126	0,126	0,126	0,126	0,126	0,126	0,126	0,126	0,126	0,126	0,126	0,126	0,126	0,126	0,126
Kurtosis		-0,036	-0,287	-0,193	0,158	-0,301	-0,427	*0,813**	-0,037	0,342	0,124	-0,085	-0,312	*-0,850**	-0,118	0,235	-0,024	0,479	-0,111
Std. Error of Kurtosis		0,251	0,251	0,251	0,251	0,251	0,251	0,251	0,251	0,251	0,251	0,251	0,251	0,251	0,251	0,251	0,251	0,251	0,251

* significant (p<0,05)