Disponível em
http://www.anpad.org.br/rac-e

RAC-Eletrônica, v. 1, n. 2, art. 8, p. 114-135,
Maio/Ago. 2007

Building a Better Measure of Business Performance

Jorge Manoel Teixeira Carneiro
Jorge Ferreira da Silva
Angela da Rocha
Luís Antônio da Rocha Dib

Resumo

Os resultados empíricos acerca dos fatores determinantes do desempenho organizacional têm sido conflitantes. Entre as possíveis razões, as diversas abordagens, freqüentemente inapropriadas, que têm sido empregadas para conceituar e medir o fenômeno podem resultar em evidências não significativas ou contraditórias. Neste trabalho, é conduzida uma revisão geral dos procedimentos de desenvolvimento de construtos e validação de modelos de mensuração, a qual é então aplicada ao caso específico do construto de desempenho organizacional. As precauções metodológicas seguidas aqui devem ser capazes de evitar as principais falhas presentes nas pesquisas conceituais e empíricas sobre o tema. Espera-se que o arcabouço conceitual, metodológico e procedimental aqui apresentado estimule outros pesquisadores a se engajar num esforço conjunto para desenvolver um modelo de mensuração, baseado em forte argumentação substantiva, rigor psicométrico e validação empírica, para o construto desempenho organizacional, que deverá ser generalizável para diversos contextos de pesquisa.

Palavras-chave: mensuração do desempenho organizacional; construto de desempenho organizacional; estratégia empresarial.

Abstract

The empirical results concerning the determinants of business performance have been conflicting. Among other possible reasons, non-significant or contradictory findings may be due to the diverse and often inappropriate approaches that have been used to conceptualize and measure the phenomenon. In this paper, a general review of construct development and measurement model validation procedures is conducted, which is then applied to the specific case of the business performance construct. The methodological cautions followed here are meant to overcome most of the flaws found in conceptual and empirical research on the subject to date. It is expected that the conceptual, methodological and procedural framework presented here will encourage other researchers to engage in a joint effort to develop a substantively based, psychometrically sound and empirically validated measurement model of the business performance construct that will be generalizable to diverse research settings.

Key words: business performance measurement; business performance construct; business strategy.

Introduction

Studies into the determinants of business performance, and their respective impacts thereof, have reached some conflicting results, as can be ascertained by a review of some of the frequently cited works in the Strategic Management literature, such as: Hansen and Wernerfelt (1989), McGahan and Porter (1997), Powell (1996), Roquebert, Phillips and Westfall (1996), Rumelt (1991), Schmalensee (1985), Wernerfelt and Montgomery (1988), among others. Such inconsistent findings may be due to poor conceptualization, operationalization and measurement of the dependent construct. Although other reasons for the inconsistent findings may be suggested (e.g., diversity in settings or in scope of analysis, diversity in the nature and number of explanatory factors employed, diversity in the types of effects modeled, diversity in conceptualization, operationalization and measurement of explanatory factors; or also the possibility that business performance might be modeled as an independent variable, instead of the usual dependent-variable approach, as suggested by March and Sutton, 1997), this paper focus specifically on the conceptualization and measurement of the business performance construct. As such, we address Boyd, Gove and Hitt's (2005) complaint that "[s]trategic management has been characterized as placing less emphasis on construct measurement than other management subfields" (p. 1) .

The objective of this paper is to present a well-substantiated framework to guide researchers through the steps of scale (or index) development and validation for the purpose of measuring an abstract construct – specifically, the business performance construct. As Peter (1981) has put it, "theories cannot develop unless there is a high degree of correspondence between abstract constructs and the procedures used to operationalize them" (p. 133).

A latent variable approach, based on structural equation modeling, will be suggested and thoroughly discussed, through which a measurement model for the construct can be conceptualized and operationalized. The framework developed here draws heavily on some previously published efforts by other scholars – either of a conceptual, empirical or methodological nature –, but is argued to have moved further by integrating and organizing related and complementary conceptual, methodological, procedural and instrumental orientations that have been scattered through several works. The main contribution of this paper is to provide researchers with a practical and readily operational set of methodological procedures to be followed in the development of a new measurement model of the business performance construct. No specific fully operational measurement model for the business performance construct will be presented in this paper, but it can be generated as the procedures advanced here are followed in future studies.

This paper is organized as follows. First, a discussion is conducted on content validity issues, followed by guidelines for the proposition of preliminary (competing) measurement structures. Third, proposals of appropriate procedures for the assessment of the psychometric properties of the measures and, fourth, several aspects of construct validity are discussed. Fifth, considerations as to generalizability and cross-national validation are carried out. Sixth, ways of building an aggregated metric and interpreting the meaning of the final score are addressed. Final reflections and a few suggestions for future studies close the paper.

Construct Conceptualization and Content Validity Considerations

The first step in the development of a new measure for a construct is to understand its basic nature and come up with an appropriate conceptual definition to guide subsequent efforts. It is then necessary to operationalize the definition by means of empirical items that sample the construct's domain. This section centers on the conceptual definition, dimensional issues, and the appropriate measurement perspective to conceive of the business performance construct. Such a discussion leads to a proposition of a generic analytical framework for the characterization of the phenomenon and to a set of recommended steps to generate an initial pool of measuring items and their subsequent screening and refinement (the terms **item** and **indicator** will be used interchangeably throughout this paper).

Conceptual Definition

A construct is an abstract theoretical (hypothetical) variable that is invented (**constructed**) to explain some phenomenon (Kerlinger, 1986, as cited in Schriesheim, Powers, Scandura, Gardiner, & Lankau, 1993), which should be given a meaning by way of a theoretical definition. A proper definition should (Hinkin, 1998): set the boundaries of coverage (what is encompassed by and what is excluded from the concept); identify the main distinct facets and, as a natural consequence, the latent variables that represent the concept (Bollen, 1989); and set an initial standard by which to select measures. Also, a good definition should make clear the extent to which the values of the construct are expected to differ across cases, settings or time (MacKenzie, 2003).

Barney (1996) contends that: “There are numerous definitions of **organizational** [business] **performance** but relatively little agreement about which definitions are ‘best’, let alone agreement about the criteria against which definitions should be judged” (p. 30, emphasis in the original). Cameron (1986) states that: “Organizational effectiveness [performance] is mainly a problem-driven construct rather than a theory-driven construct” (p. 541). In fact, the conceptual definition of business performance should be oriented by the specific objectives of the firm for its business units (e.g., short-term economic or market performance *versus* longer-term strategic objectives), so that a coherent set of measurement items can be drawn. Comparison of results across firms (or business units) can only meaningfully be done when the same conceptual definition applies to all of them.

The fact that there still seems to be no clear, widely accepted, definition of business performance on which to base a set of indicators should be explicitly addressed in research.

Dimensional Issues and Measurement Perspective

It is important to resolve whether the construct seems to be uni- or multi-dimensional (one single main facet *vs*. multiple facets). However, whether a construct ought to be viewed as unidimensional or multidimensional depends on the level of abstraction used to define it (Jarvis, MacKenzie, & Podsakoff, 2003), since one can look at each facet as a separate construct, but at a more abstract level all facets are integral parts of the overall construct.

The researcher should specify the (expected) relations between the measures (also called items or indicators) and the latent variables. Venkatraman and Ramanujam (1986) advise that: “…researchers should collect data on indicators of business performance either using an a priori classification which recognizes the dimensionality issue, or they should explicitly test the dimensionality of their conceptualization of business performance” (p. 807).

If there seems to be a single main facet, a reflective measurement perspective is implicit. If there seem to be two or more facets, it is necessary to determine, based on substantive sense, whether a reflective or a formative perspective seems to be more plausible (Bollen & Lennox, 1991; Diamantopoulos, 1999). In a reflective measurement perspective, the (observed) items are considered or assumed to be effects of an underlying latent construct, whereas in a formative measurement perspective, the items are assumed to **cause** a latent construct.

Jarvis et al. (2003) suggest that a construct should be modeled in a formative perspective if its measures satisfy (most of) the following conditions:

. Hypothesized direction of **causality**: the indicators are viewed as characteristics of the construct (phenomenon) rather than manifestations of it; changes in the indicators are expected to cause changes in the construct; changes in the construct are not necessarily expected to **cause** (be associated with) change in (all) the indicators;

. Interchangeability of indicators: the indicators do not necessarily have the same or similar content, do not necessarily share the same theme and do not necessarily capture the same or similar aspects of the construct; eliminating an indicator may alter the conceptual domain of the construct (that is, the remaining indicators as a set would not represent well the domain of the construct);

. Covariation among the indicators: a change in the value of one of the indicators is not necessarily expected to be associated with a change in all the other indicators; and

. Nomological net of the construct indicators: The indicators are not expected to have the same antecedents and consequences.

Furthermore, the degree of complexity of the construct should also be discussed. Multiple levels of abstraction might give a clearer picture of the phenomenon, but choosing too high a model order may add complexity without a corresponding increase in explanatory power. Besides, not only would it make the data-collecting instrument too long, but it may also, given managerial limitations to come up with information in a very disaggregated form, be difficult to elicit details that are too fine-grained about the specific aspects of business performance. Therefore, some trade-off has to be sought.

It is critical that the appropriate measurement perspective be explicitly chosen because it will have relevant implications on how to judge the quality of the new measurement model. Also, model misspecification can result in either Type I or Type II error (Jarvis et al., 2003). That is, by inadvertently inverting the **cause-effect** relationship between business performance and its indicators one may mistakenly conclude that a relationship (between business performance and some of its supposed determinants) exists when in fact it does not (Type I error), or else fail to detect a relationship when it in fact exists (Type II error).

Presentation of a Generic Analytical Framework

Carneiro, Silva, Rocha and Hemais (2005) advanced a generic analytical framework (Table 1) for the characterization of the business performance construct – the authors admittedly drew heavily from, although they also revised and extended, previously published work, especially that of Matthyssens and Pauwels (1996) and Katsikeas, Leoniodu and Morgan (2000). Such a generic scheme seems to adequately represent the rather complex domain of the phenomenon and could serve as a starting point for a guiding yardstick to elicit appropriate measures. Each form to measure the construct has its own advantages and limitations, which have been thoroughly discussed in the literature and will not be revised here.

Table 1: Generic Analytical Framework for the Characterization of the Business Performance Construct

Conceptual Aspects				Methodological Decisions		
Stakeholder viewpoint	**Classes of measures**	**Frame of reference**	**Temporal orientation**	**Unit of analysis**	**Mode of assessment**	**Indicators structure**
▪ stockholders ▪ clients ▪ employees ▪ managers ▪ debt holders ▪ suppliers ▪ channels ▪ business partners ▪ local community ▪ governments	▪ economic ▪ market ▪ internal business processes ▪ innovation and learning ▪ strategic ▪ social ▪ environmental ▪ behavioral / situational ▪ overall	▪ absolute ▪ relative –main competitors' average –benchmark –other SBU's in the firm –pre-set goals	▪ static –recent past –future expectations ▪ dynamic –change in recent past –expected change for the future	▪ country / region ▪ industry ▪ whole firm ▪ division / SBU (strategic business unit) ▪ product-market venture	▪ objective –secondary sources –self-reported ▪ subjective (primary sources) –self-evaluation –evaluation by competitors –evaluation by experts	▪ independent indicators –single –multiple ▪ composite scales –reflective –formative

Source: Adapted from Carneiro et al., 2005.

A look at recent empirical research and even classical works shows that most studies have used very simple, therefore, inappropriate conceptual definitions and operationalizations, which do not adequately capture the multifaceted nature of the phenomenon – most of them employing only one or very few indicators of business performance, usually taken as independent, i.e., not arranged into a latent variable representation. Moreover, an enormous diversity of measures (e.g., accounting-based, market-based and perceptual measures) and measurement schemes has been employed (see Venkatraman & Ramanujam [1986] and Tehrani & Noubary [2005] for a partial review of such practices). Such a conclusion reinforces how important it is to develop and validate a more robust, consistent and generalizable measurement model for the business performance construct.

Item Generation and Purification

In addition to the conceptual definition, which covers philosophical aspects, one also needs an operational definition for the construct. Complex concepts cannot be directly measured and must instead be estimated from multiple indicators. So it is necessary to generate a set of items thought to represent the construct and demonstrate that such empirical items are logically and theoretically connected to it (Nunnally, 1978). The set of items should fully represent all key aspects of the theoretical domain of interest (Bollen, 1989; MacKenzie, 2003), but care has to be taken in order not to contaminate the set of items by the inclusion of other aspects that are not part of the conceptual domain (Churchill, 1979).

A thorough literature review coupled with exploratory interviews with practitioners and academicians – both carefully guided by solid theoretical foundations – are the first steps towards generating an initial pool of items. As a mere example, exploratory interviews could include such questions as: a) What does [the construct] mean to you? b) What are the main aspects that one ought to consider when measuring [the construct]? c) Give examples of measures you would use to assess the [degree / amount of the construct]; d) Give examples of a [firm] that you think would score high on [the construct] and another [firm] that would score low. In what aspects do [the construct] values differ between these two [firms]? Then, the researcher would analyze the items thus elicited and conduct a first screening by removing items that seem inconsistent and by collapsing together items that seem to have the same meaning, albeit having been worded differently. It is important to refine the wording in the descriptions of items so that respondents and researchers will interpret the questions in a similar manner.

Table 2 presents some possible business performance indicators that serve as an illustration of how the several conceptual aspects advanced in Table 1 can be operationalized.

Table 2: Suggestion of Possible Indicators of Business Performance

Indicators		Conceptual aspects covered
E_1	satisfaction with [SBU's] revenues in last three years	economic, absolute, static (recent past)
E_2	revenues growth of the focal SBU *vis-à-vis* revenues of other SBU's of the firm in last three years	economic, relative (to other SBU's in the firm), dynamic (recent past)
E_3	expected SBU profitability for next three years	economic, absolute, static (near future)
M_1	SBU volume *vis-à-vis* competitors in last three years	market, relative (to competitors), static (recent past)
M_2	expected volume of the focal SBU vis-à-vis volume of other SBU's of the firm for next three years	market, relative (to other SBU's in the firm), static (near future)
M_3	SBU volume growth in last three years	market, absolute, dynamic (recent past)
O_1	overall SBU results in last three years	overall, absolute, static (recent past)
O_2	expected overall SBU results for next three years	overall, absolute, static (near future)

Note: these indicators were extracted and adapted from a list of over one hundred business performance indicators found in several empirical works (the full list is available from the authors upon request).

The next step is the purification phase. Although there does not seem to be any single general rule or technique for ensuring content validity, Anderson and Gerbing (1991), MacKenzie, Podsakoff and Fetter (1991), Schriesheim et al. (1993), and MacKenzie (2003) have provided some suggestions. The basic idea is to prepare a new questionnaire – with the whole set of items in the lines and with definitions of the possible dimensions (aspects or facets of the construct) devised by the researcher in the columns – and ask respondents to associate items with dimensions or rate the similarity of items or rate the extent to which they believe each item corresponds to each specific dimension definition. The facets shown to these judges should be both mutually exclusive and exhaustive of the content of the domain of the focal construct, including a **none of the above** or **other** category. Straub, Boudreau and Gefen (2004) recommend that questions relating to the several dimensions of the construct should be randomized in order to prevent respondents from sensing the inherent constructs via the ordering of the questions and responding accordingly. MacKenzie (2003) and DeVellis (1991) provide additional guidelines for the preparation of this pilot questionnaire.

As for the appropriate number of items to consider in the questionnaire, Hinkin (1998) suggests one should generate around twice as many items (per trait or dimension) as are expected to be retained after this purification phase. The important point is to come up with the minimum number of items that seem to adequately tap the domain of interest, instead of blindly trying to meet any rule of thumb as to the **adequate** number of indicators. Avoiding too many items will also minimize the response bias caused by fatigue or boredom. If model identification requirements or validation needs (detailed in the sections ahead) demand additional items, these should be included in the questionnaire. Factor analytic techniques, multidimensional scaling or hierarchical clustering could then be used to determine the content grouping of the items.

This whole procedure would expectedly lead to the identification of (i) further items and dimensions deemed necessary to represent the construct, (ii) which essential aspects of the conceptual domain are tapped by each of the items, and (iii) inconsistent items – those that tap some (confounding) constructs which are not part of the focal construct's conceptual domain.

Item Scoring

Once the researcher is satisfied with the variety of items that has resulted from the purification steps just described, he/she has to decide about the appropriate scoring (scaling) of the items for use in subsequent phases of measurement model building. The response scale should be capable of generating sufficient variance among respondents for subsequent statistical analyses, so one should accurately benchmark the response range before building up the response scale (Bass, Cascio, & O'Connor, 1974). Besides, item scaling and anchor words should be such as to avoid leniency (tendency to rate higher than deserved) and increase precision (agreement among raters) (Menezes & Elbert, 1979).

The number of response alternatives should be dictated by the needs of each study and should not be too large or too small so that distinctions that are meaningful to respondents should be represented (Menezes & Elbert, 1979). Five-point Likert scales and seven-point semantic differential scales would fit the requirements of most studies. In fact, coefficient alpha reliability with Likert scales has been shown to increase up to the use of five points and then it levels off (Lissitz & Green, 1975). Cox (1980) additionally advises that "an odd rather than an even number of responses is preferable under circumstances in which the respondent can legitimately adopt a neutral position" (p. 420).

Generation of Preliminary Theoretically-Based Measurement Structures

Model Identification Requirements

Some model specifications would render the model unidentifiable from a structural equation modeling standpoint. Identification means that the model contains enough information in order for a single **best** solution (for a set of structural equations parameters) to be estimated (Hair, Black, Babin, Anderson, & Tatham 2005). The main (measurement) parameters to be estimated are the links between the indicators (observed variables) and their respective latent variables as well as the links between latent variables in the measurement model and also the error terms. And the information used for such estimation refers to the number of variances and co-variances among indicators, so there should be a sufficiently large number of indicators in order to provide enough information.

MacCallum and Browne (1993) argue that: "Models that are not identified should not be used in practice because they contain parameters whose values are arbitrary" (p. 537). In terms of a reflective indicator measurement model, at least three effect indicators are necessary in order to make it identifiable (Long, 1983, as cited in Diamantopoulos & Winklhofer, 2001). Furthermore, taken in isolation, a formative indicator measurement model is statistically underidentified (Bollen & Lennox, 1991) because of indeterminacies associated with the scale of measurement of the construct (which, in principle, is arbitrary) and with the construct error term (MacCallum & Browne, 1993) – and it can be estimated only if it is placed within a larger model that incorporates consequences (that is, effects) of the latent variable in question (Bollen, 1989).

The scaling issue can be resolved by arbitrarily constraining one of the paths from one of the indicators to be equal to some nonzero value, usually one (thus equating the scale of measurement of this construct to that of the respective indicator) or by constraining the residual error variance of the construct to be equal to some nonzero value, usually 1.0 (MacCallum & Browne, 1993). As for the indeterminacy of the construct level error term, it is necessary that the formative construct should emit paths to (a) at least two unrelated latent constructs measured by reflective indicators, or (b) at least two (theoretically appropriate) reflective indicators of the construct itself, or (c) one reflective indicator and one latent construct measured with reflective indicators (Diamantopoulos & Winklhofer, 2001; Jarvis et al., 2003; MacCallum & Browne, 1993). Otherwise, one can fix the error term to zero or equate it with the residual variance associated with the construct it is hypothesized to influence, but such procedures may not be theoretically appropriate (MacCallum & Browne, 1993).

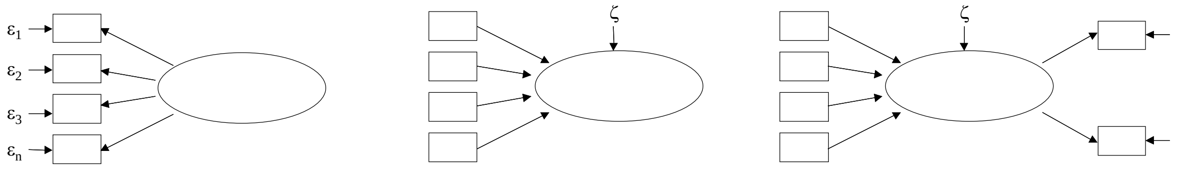
All in all, if identification requirements demand additional variables, the researcher should anticipate such needs for inclusion in the questionnaire concerning the indicators.

Theoretically-based Model Structures

At this stage, the researcher can conceive of possible measurement structures, based on substantive arguments, which make explicit the possible relationships between indicators and construct as well as the expected degree of complexity (existence and relationship of sub-constructs (dimensions) to a higher-order construct).

Figure 1 presents three generic unidimensional specifications, including a MIMIC (multiple indicators multiple causes) model. Figure 2 illustrates three generic multidimensional first-order specifications, while Figure 3 illustrates five generic multidimensional second-order specifications. Note that these 11 specifications are not intended to provide an exhaustive account of all possible measurement models, but rather to serve as indicative of the main possibilities (see Edwards & Bagozzi [2000] for other possible alternatives.)

Figure 1: Generic Unidimensional Specifications



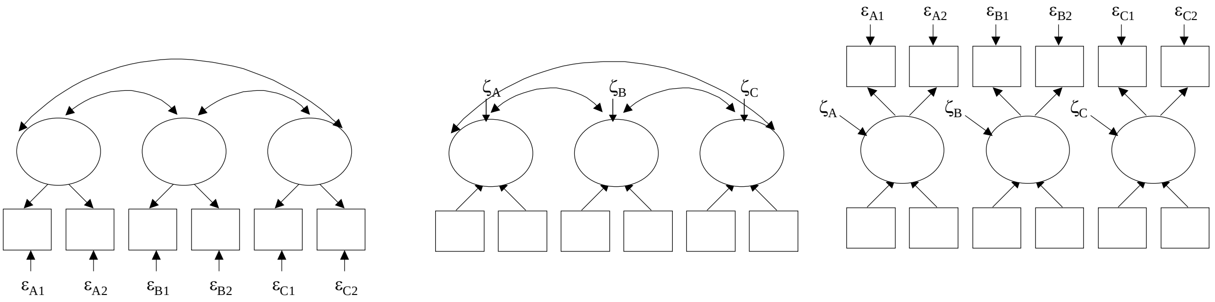
a) Unidimensional reflective representation

b) Unidimensional formative representation

c) Unidimensional MIMIC representation

Note: measurement error in the formative indicators has been omitted, as is usually assumed (cf. MacCallum & Browne, 1993), though in fact untenable in most situations (Edwards & Bagozzi, 2000).

Figure 2: Generic Multidimensional First-order Specifications



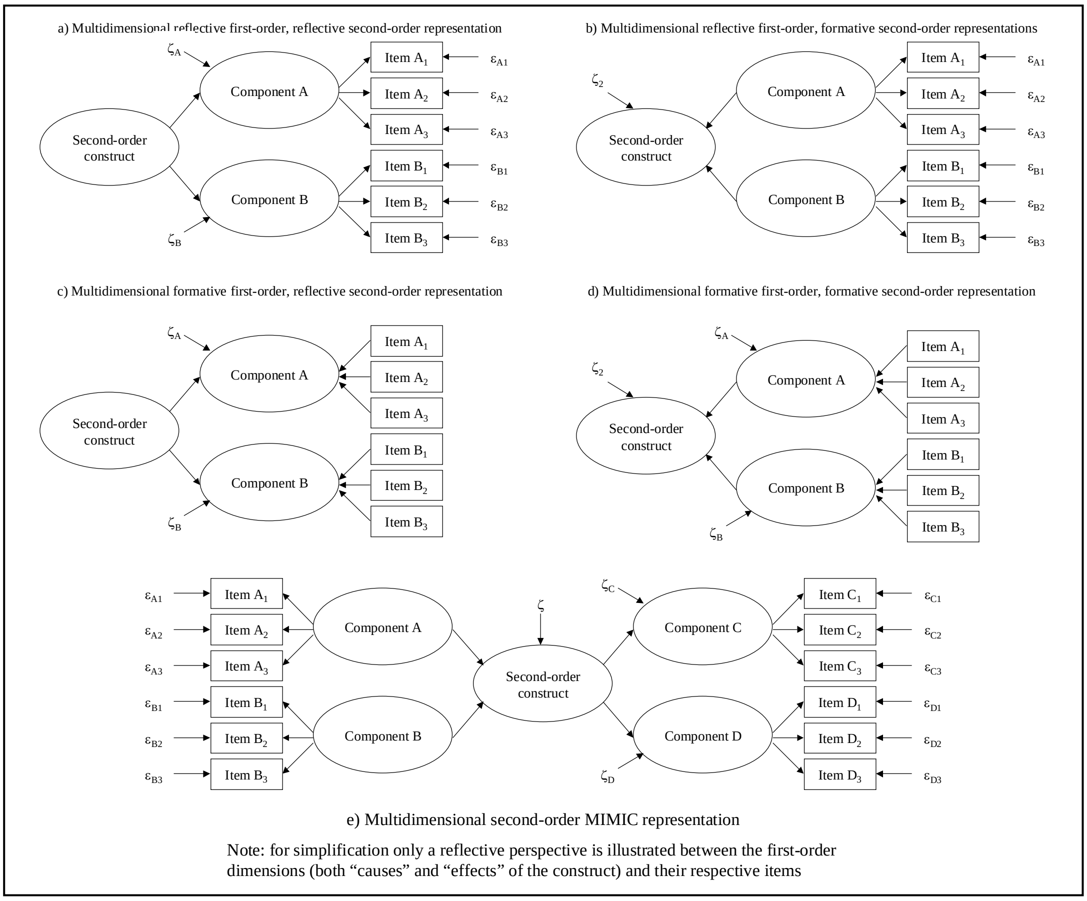
a) Multidimensional first-order reflective representation

b) Multidimensional first-order formative representation

c) Multidimensional first-order MIMIC representation

Source: Adapted and expanded from Rindskopf and Rose (1988).

Figure 3: Generic Multidimensional Second-order Specifications



a) Multidimensional reflective first-order, reflective second-order representation

b) Multidimensional reflective first-order, formative second-order representations

c) Multidimensional formative first-order, reflective second-order representation

d) Multidimensional formative first-order, formative second-order representation

e) Multidimensional second-order MIMIC representation

Note: for simplification only a reflective perspective is illustrated between the first-order dimensions (both "causes" and "effects" of the construct) and their respective items

Each study shall determine which specific aspects of the business performance phenomenon are more relevant, since it is not practically feasible to collect information on all the aspects outlined Table 1. Some alternative possible measurement structures will be advanced here, making explicit (i) the order complexity, i.e., the existence and relationship of dimensions to a higher-order construct, (ii) the underlying structure (association and measurement perspective) between indicators (and dimensions) and construct; and (iii) the number of measuring indicators. They will serve as the basis for the construction of alternative operational measuring instruments.

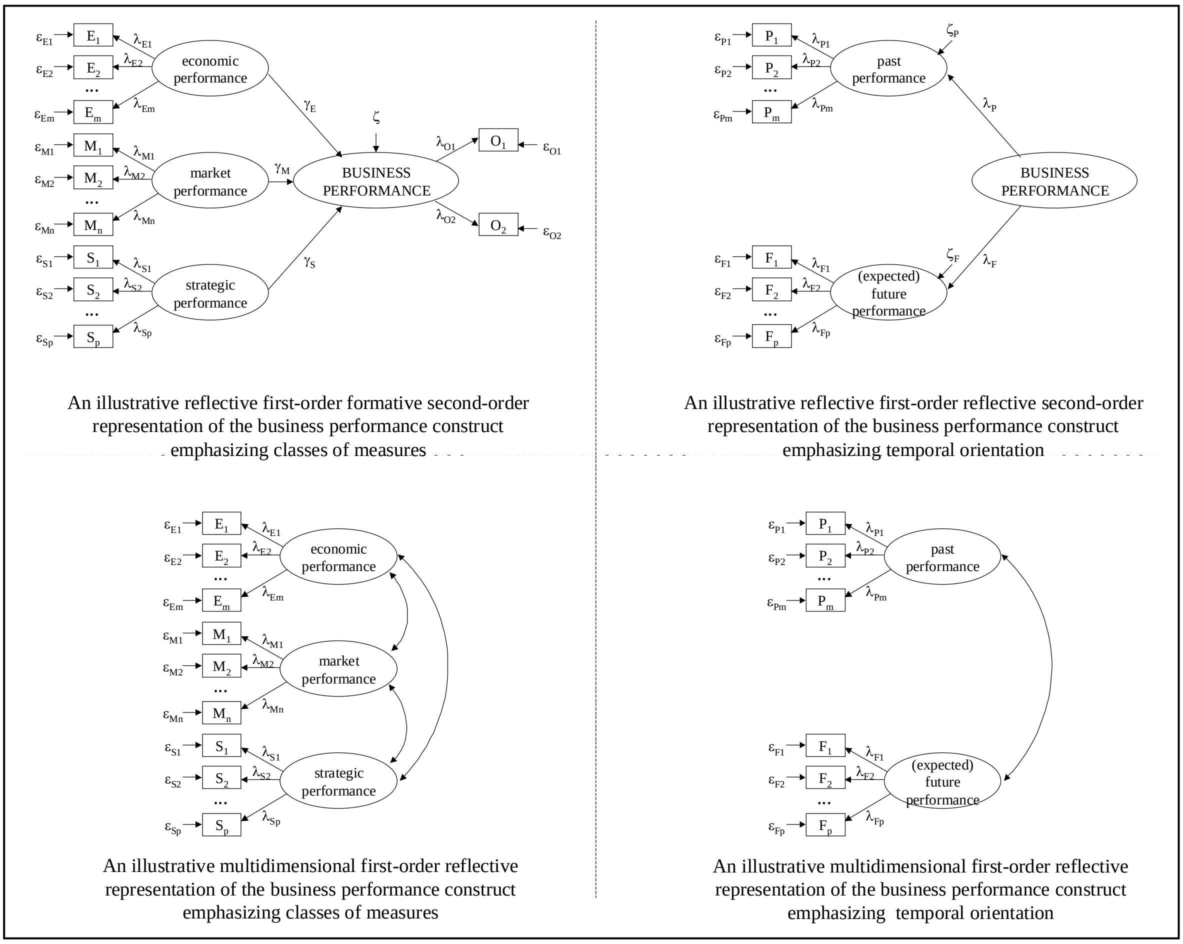
At a basic level, it is reasonable to think of indicators as reflective, that is, a set of economic indicators would be expected to co-vary together, as would another set of market indicators (as representations respectively of the economic and market facets of business performance), although the two sets, as wholes, might not show strong correlation. However, at a more abstract level, it seems appropriate to conceptualize the performance construct as being determined by (instead of determining) its indicators (or dimensions) (Diamantopoulos, 1999).

Therefore, specifications (a) and (b) in Figure 1 seem to be too simple to represent the multifaceted nature of the business performance construct, but should not immediately be ruled out, but rather kept for later analysis and comparison. Except for specification 3-c, which does not seem to make theoretical sense, and specification 3-e, which may be too complex, all other specifications should also be kept for assessment of their psychometric properties and evaluation of construct validity; however, specifications 3-b and 3-d would render the model statistically unidentifiable from a structural equation modeling standpoint (fewer unique values in the covariance matrix than parameters to be estimated), thus demanding additional (reflective) indicators.

So, a MIMIC representation could be used in order to make the model identifiable (Bollen, 1989) – in fact reflective indicators of the construct can be added to an otherwise formative model –, but substantive reasoning will have to be employed because a MIMIC model can have several conceptual interpretations (Jarvis et al., 2003), which are indistinguishable from a mathematical standpoint (since they all produce identical estimates of the relationships between the measures and the constructs). Three possible interpretations are: a single construct with *n* formative and *m* reflective indicators; *n* exogenous variables influencing a single endogenous construct measured with *m* reflective indicators; and a formatively measured construct that influences *m* manifest measures of (one or more) different constructs.

Given such considerations, and after an initial pool of items is generated (see ahead), some illustrative competing measurement models (say, six to eight) could be advanced (see Figure 4).

Figure 4: Illustrative Representations of the Business Performance Construct



An illustrative reflective first-order formative second-order representation of the business performance construct emphasizing classes of measures

An illustrative reflective first-order reflective second-order representation of the business performance construct emphasizing temporal orientation

An illustrative multidimensional first-order reflective representation of the business performance construct emphasizing classes of measures

An illustrative multidimensional first-order reflective representation of the business performance construct emphasizing temporal orientation

Assessment of the Psychometric Properties and Refinement of the Preliminary Measurement Models

A questionnaire can now be administered to managers and data can be collected concerning the performance indicators of real firms (or SBU’s or product-market ventures).

Indicator values can be obtained from **objective** sources or in a **subjective** fashion. **Objective** data may come from secondary sources (e.g., company’s annual reports or press releases) or it may be reported by executives. In this case, the respondent is not asked to express a personal opinion, but to provide information that, supposedly, would be reported in the same way no matter who provides it. It is worth noting that some of the most prominent works that have sought to explain the components of variance of performance (e.g., McGahan & Porter, 1997; Roquebert et al., 1996; Rumelt, 1991; Schmalensee, 1985) were based on objective data from secondary sources, but the **objectivity** and comparability of such type of data has often been questioned given differing accounting practices and valuation methods of intangible assets as well as managerial manipulation (Barney, 1996). On the other hand, subjective measures (see, for example, Powell, 1996) capture the respondent's opinion or perception of performance and are especially important when objective measures are missing, which is usually the case in small or privately-held firms or in newly-launched operations. Since the way executives perceive and judge performance is a key driving force of their future actions (Dess & Robinson, 1984), perceptual measures should be obtained. By asking for the respondents' opinion about perceived success or satisfaction with achievement of objectives, the researcher will be able to get a measure that somehow incorporates and consolidates all others.

The questionnaire should contain about twice as many items (for each of the dimensions) as are expected to be retained after this new purification phase (e.g., if one expects to have four items per dimension, start off with eight, cf. Hinkin, 1998). Supposing that most of the possible competing measurement models that the researcher may conceive of would contemplate three dimensions (e.g., economic plus market plus strategic business performance) or four dimensions (past plus expected future plus past variation plus expected future variation of business performance) and considering that between four and five items per dimension would be a good final target, one ought to expect to retain around 14-18 items in the final instrument. Considering the additional need for some three overall indicators (for identification requirements and criterion-related validity checks), this would add up to 35-40 items (that is, around twice as many) in the questionnaire. Given a 10:1 target ratio of cases to items (Hair et al., 2005), one ought to expect to work with a sample of around 400 firms – which is quite large and, moreover, not all of the firms may have published data, which would force the researcher to rely on perceptual data.

Now that some preliminary measurement models have been designed, it is time to assess their **quality** or **satisfactoriness**, that is, how well the items that survived the content validity phase confirm expectations as to the desirable psychometric properties, given the proposed model structures. Care should also be taken in order to understand the different approaches to be followed when assessing the psychometric properties of reflectively measured *versus* formatively measured constructs.

No blind rule to clean the model and assess its satisfactoriness should be universally applied, since the right decision would depend on the specific measurement perspective (that is, reflective versus formative). Conflicting results between substantive reasoning and obtained empirical relationships should lead to further investigation of the construct's content.

In fact, the decision about which indicators should be assigned to which facets of the business performance construct should be dictated by substantive reasoning. Moreover, the specification of the relationships (i.e., reflective vs. formative) of the indicators to their respective facets should be based on substantive reasoning about whether a given indicator seems to be determined by the construct or, instead, seems to **cause** it. If a higher-order structure seems to better express the construct's nature, similar considerations ought to be applied between first- and second-order dimensions.

Guidelines to Judge the Quality of the Measurement Model in a Reflective Structure

Exploratory factor analysis – to be used to cluster items into dimensions (factors) –might lead to more or fewer dimensions than might have been considered in the previous phase. Furthermore, one ought to investigate whether there is some pattern of correlation among the dimensions thus derived that might offer support for higher-order dimensions hypotheses (Marsh & Hocevar, 1985). Rotation can ease interpretation. Orthogonal rotation would maintain dimension independence (assuming traits are uncorrelated), but oblique rotation may provide richer insights and avoid distorted factor loadings and incorrect conclusions about the number of factors (Bagozzi & Phillips, 1982) in case the judges experience difficulties with the discrimination of the rating dimensions.

Good reflective indicators should exhibit high internal consistency – which includes both unidimensionality (extent to which the items reflect one single underlying trait) and reliability (accuracy or precision of the measuring instrument) (Bollen & Lennox, 1991). High internal consistency is a necessary but not sufficient condition to ensure construct validity because the source of common (systematic) variance may be due to contamination (a third common source) or the items may not exhaustively represent the construct.

Unidimensionality should be assessed for each separate subscale, each one measuring one of the distinct dimensions of the construct (Peter, 1981). In order to get unidimensional measurement, the researcher should delete items (i) that do not load high (e.g., less than .30) on any (rotated) dimension or (ii) whose correlation with all other items in the same dimension (inter-item correlation) or with the composite set of other items (minus itself) in the same dimension (item-to-total correlation) is small (less than .30 and .50, respectively) or (iii) that load high (more than .30) on more than one dimension (Epstein, 1983; Kim & Mueller, 1978). For the sake of parsimony, those items should also be removed (iv) whose loading is not statically significant (even if it is **large**) or (v) which correlate too high (over .90) with others in the same dimension (redundancy, according to Briggs & Cheeks, 1986). However, one should remember that if errors of measurement are correlated, inter-item correlation may be inflated or deflated (Smith, 1999), so appropriate covariance algebra should be used to derive the composition of the correlation (Bollen & Lennox, 1991).

Reliability is a necessary but not sufficient condition for construct validity because "The observables may all relate to the same construct but that does not prove that they relate to the specific construct […]" (Churchill, 1979, p. 72). Furthermore, once demonstrated for a given sample, the reliability of a scale cannot simply be assumed to hold universally, since it is a situational indicator of the effectiveness of the measurement instrument (Nunnally, 1978) and must be demonstrated a posteriori for every sample to which it is administered (Schrieshiem et al., 1993). There are several methods for estimating reliability (Bollen, 1989), for example: test-retest, alternative forms, split-halves, coefficient alpha, composite reliability index – all of which have their own advantages and limitations. Coefficient alpha (also known as Cronbach's alpha) is probably the most frequently employed reliability measure in empirical research. Compared to other methods, alpha makes the least restrictive assumptions: no temporal stability requirements (as in test-retest and in alternative forms), no need for parallel measures (i.e., equal scores and equal error variances and zero correlation between the errors in the two different points in time), possibility of memory effects (as in test-retest) is remote, no problem of grouping indicators (as in split-haves). There is also the composite (or construct) reliability index (Hair, Anderson, Tatham, & Black, 1998), which, unlike coefficient alpha, has the advantage that it does not assume that indicators have equal factor loadings and error variances, but rather takes into account differences among the indicators (Styles, 1998). Alpha reliability greater than 0.7 and composite reliability greater than 0.6 are usually considered acceptable. Also, the researcher should verify whether average variance extracted high (over .50, which would mean that more than 50% of the variance of the indicators would be accounted for by their respective latent sub-construct, Hair et al., 1998).

The researcher should try to remove items that, once removed, do not significantly negatively affect reliability or that, once removed, improve reliability (unless such items are deemed relevant due to theoretical considerations), but such a procedure is justified only after the unidimensionality of the scales has been established (Gerbing & Anderson, 1988; Hinkin, 1998). Carmines and Zeller (1979) point out that the addition of items makes progressively less of an impact on reliability. The analysis should be repeated until one reaches a measurement model that is satisfactory in terms of reliability and that explains a high percentage of total item variance and is reasonably parsimonious at the same time.

Also, confirmatory factor analysis (CFA) can be used to assess overall model fit by means of the chi-square (χ^2) statistic (a χ^2 two or three times as large as the value of the degrees of freedom is acceptable – Carmines & Mcver, 1981, as cited in Hinkin, 1998). Small *p*-values indicate that the hypothesized structure is not confirmed by the sample data (Hughes, Price, & Marrs, 1986). It is important to assess the overall statistical significance of the whole model as well as that of items loadings on factors (in this case, dimensions of business performance) in order to get a stricter interpretation of unidimensionality (Hinkin, 1998). As for measurement errors (of indicators), the researcher can allow them to correlate (which tends to improve model fit), but such an assumption should be based on substantive reasoning. Moreover, such a correlation should be assessed against the possibility of the existence of a true underlying structure that would link the two indicators and that should, as such, be explicitly incorporated into the model (Hughes et al., 1986). Modification indices (provided by CFA) can be employed to respecify the model, but the modification should be theoretically acceptable and, afterwards, reliability checks should be conducted again.

Guidelines to Judge the Quality of the Measurement Model in a Formative Structure

In the case of formatively measured constructs, the magnitude of indicator correlations is not explained by the measurement model, so internal consistency (in the conventional sense) is not a pre-requisite for valid measures (Bollen & Lennox, 1991). Moreover, one should resist the temptation to delete items as a means of improving reliability, since dropping items might affect content validity of a formatively measured construct (MacKenzie, 2003). Rather, **good** formative indicators should be parsimonious, have statistically significant coefficients (loadings), lead to good overall model fit and to high explained variance.

For the sake of parsimony and also for ease of explanation, there should be removed those items which show high multicollinearity with others in the same dimension (sub-construct), e.g., tolerance less than .30). Besides, the revised model should be estimated by means of CFA – including all **dimensions** (lower- as well as higher-order) and respective remaining items as well as additional (reflective) indicators needed for model identification – and items whose coefficients are not statistically significant should be removed. Items should be deleted one a time (starting with the lowest *t* value) and the model should be estimated again, until all remaining coefficients are statistically significant.

An item that meets the aforementioned conditions for deletion may nonetheless be kept if (i) it correlates high and with the right sign with some appropriately chosen external criterion (e.g., some overall assessment of the construct, which constitutes a concurrent validity check) or (ii) its removal would result in significant deterioration either in fit (a significant $\Delta\text{-}\chi^2$) or in explained variance of the latent variable or (iii) its removal would jeopardize breadth of coverage of the construct's content.

Overall model fit should be assessed and modification indices can be employed to respecify the model. However, since a census of indicators is necessary to correctly represent a construct in a formative structure (Diamantopoulos, 1999), "indicator elimination […] should not be divorced from conceptual considerations when a formative measurement model is involved" (Diamantopoulos & Winklhofer, 2001, p. 273).

Comparative Assessment of Competing Models

The researcher should now compare models against one another. Other competing (alternative) measurement models (equivalent models can also be conceived of), composed of the items retained in the previous phase, can also be advanced. Their construction being driven by substantive reasoning, competing models can vary in terms of: number of indicators, underlying structure between indicators and dimensions, degree of complexity (number of abstraction levels), the nature and direction of the relationships between the construct and the sub-constructs (dimensions), the nature and direction of the relationships among items and sub-constructs.

Models should be ranked in terms of their overall **quality** or **satisfactoriness** as operational representations of the construct, based on theoretical and empirical considerations. They should go through checks of psychometric properties and construct validity. Also, fit indices that take parsimony into consideration (Hair et al., 1998) can be used to compare models. This model generation plus purification method can be repeated several times, until the researcher feels he/she has reached models that seem substantively and empirically sound. However, in order to avoid capitalization on chance, it is advisable to use a new sample in order to judge the **satisfactoriness** of the measurement model every time it is respecified.

Possible empirical differences observed among the alternative measurement models tested – for example, in terms of parsimony (number of indicators), content (specific indicators retained, specific dimensions resulting from the arrangement of the indicators, as well as the proportional distribution of indicators across dimensions), and construct validity – should be interpreted and understood before one chooses one model to the detriment of others (cf. Diamantopoulos & Siguaw, 2006). When in doubt, the researcher is advised to keep more than one model and conduct additional research with new samples in diverse settings.

VALIDATION OF THE MEASUREMENT MODELS

Construct validity represents "the correspondence between the conceptual definition of a construct and the operational procedure to measure or manipulate the construct" (Schwab, 1980, p. 5) and includes unidimensionality, reliability and validity itself. As Walker, Olson, Celsi, and Chow (1992) have put it, "[we] lack the necessary criteria to unambiguously determine the correctness, reality, or truth of a construct" (p. 188). This means that construct validity cannot be assessed directly, but only inferred (Peter, 1981), but the researcher should strive to develop strong support for it (Bollen, 1989). Although the previous steps for the development of measures for a construct and the assessment of its psychometric properties already tend to lead to construct validity, further evidence should be sought, e.g., criterion-related validity (which includes concurrent and predictive validity), nomological validity, convergent validity and discriminant validity.

Criterion-related validity is the degree to which a measure exhibits empirical relationships which are consonant with the theory underlying the construct (Bollen, 1989). The criterion variable (to which the newly developed measure is to be compared) has to be an accepted standard (having been previously determined to exhibit at least some degree of content validity, reliability and construct validity, *cf.* Bollen, 1989) – however, such a standard does not always exist, as seems to be the case with business performance. Furthermore, such a criterion variable should be measured in a reflective fashion and also there should be solid theoretical reasons to justify this expected relationship (Diamantopoulos & Winklhofer, 2001). Criterion-related validity gains particular importance in formative measurement because "associations of formative measures with their construct are determined primarily by the relationships between these measures and measures of [reflectively measured] constructs that are dependent on the construct of interest" (Edwards & Bagozzi, 2000, p. 159).

Concurrent validity is a sub-type of criterion-related validity when the criterion variable exists at the same time as the newly developed measure (Bollen, 1989). Smith (1999) reports that researchers can demonstrate concurrent validity by regressing factors derived from factor analysis onto overall assessments of the construct, rated on a separate scale. Diamantopoulos and Winklhofer (2001) express a similar argument: "[Another] possibility is to use as an external criterion a global item that summarizes the essence of the construct that the index purports to measure" (p. 272). So, the researcher can devise an additional statement that somehow summarizes the construct – this would be an overall indicator related to satisfaction with performance, attainment of objectives, degree of perceived success attained etc. – and test how the other indicators correlate with it. Also, the degree of association with known cases (e.g., firms or SBU's that represent clear and undisputed instances of **successes** or of **failures**) could be used to test for concurrent validity. In the case of a formative perspective of measurement, validation can also be accomplished by including some reflective indicators and estimating a MIMIC model, which evaluates the proposed indicators as a set, taking account of their interrelationships (Diamantopoulos & Winklhofer, 2001).

Predictive validity, which is another sub-type of criterion-related validity, can be demonstrated by the ability of the new measure to predict responses to questions or intentions of future behavior (Smith, 1999). In the case of business performance measurement, some possibilities are: 'Would you recommend that your firm keep investing efforts in this firm (or SBU or product-market venture)?'; or 'If you could go back in time and know how the facts would unfold, would you have recommended that the same (or higher) amount of financial and managerial resources be invested in this firm?'.

Such tests could in fact be conducted for different versions of measurement model (to see which seem to show greater concurrent or predictive validity) and also for different versions of the overall scores and of the statements of future behavior. Then, one could evaluate concurrent and predictive ability in terms of dimensions of the construct involved and also in terms of control groups (e.g., large vs. small firms; developed vs. developing countries, products vs. Services, etc.), i.e., which dimensions seem to have a greater concurrent validity and predictive ability than others across which groups.

Nomological validity refers to the degree to which predictions from a theoretical network containing the concept are confirmed or the extent to which a construct relates to other constructs in a predictable manner (Venkatraman, 1989). According to Venkatraman and Grant (1986), "[…] predictive validity entails the relationship of a measure of a construct to a single antecedent or consequent; nomological validity involves many constructs in a complex system" (p. 82). The researcher can make use of another construct that is expected to be affected by the one under development. If the estimated model shows a significant path with the expected sign between the two constructs, this would serve as additional evidence of validity. But there the risk of circular reasoning here, since this last condition practically implies that the hypotheses relating the two constructs would be irrefutable because refutation might be attributable to construct validity problems (Carmines & Zeller, 1979).

Since methods to assess business performance may not always be fully reliable, it is recommended that the researcher also check for convergent validity, that is, the degree to which multiple attempts to measure the same concept with (maximally) dissimilar methods are in agreement (Campbell & Fiske, 1959). Different methods can include: objective *vs*. subjective data, interviews, questionnaires, archival data, multiple managers in different key functions, published secondary data, expert opinion, and use of different types of scales (Venkatraman & Grant, 1986) or different respondents outside the firm. Especially in the case of objective vs. subjective data comparison, it may be advisable to rescale objective data (e.g., into a 5- or 7-point Likert-like scale) so that it can be directly compared with subjective data (as suggested by Tehrani & Noubary, 2005). If the correlation between measures is "significantly different from zero and sufficiently large" (Campbell & Fiske, 1959, p. 82), this provides evidence of convergent validity. Researchers can also check whether all items load significantly on their respective (hypothesized) sub-scales and (*cf.* Straub et al., 2004) if they show high and significant correlations with one another (if a reflective perspective is employed) – this is akin to the test of unidimensionality, except that it should be run only with items measured by dissimilar methods. When there are already established and accepted measures for the focal construct (e.g., drawn from a literature review), they could be used to check for convergent validity with the new measure being developed (Hinkin, 1998). However, in the case of business performance, there does not seem to be a clearly acceptable standard against which to compare the new measure.

Furthermore, the new construct should be shown to exhibit discriminant validity, which is the degree to which measures of distinct concepts differ (Venkatraman & Grant, 1986) – that is, it should be demonstrated that the measure does not correlate very high with another measure from which it should differ (Peter, 1981), albeit being related to it. Since the frequently cited MTMM (multitrait multimethod) approach (Campbell & Fiske, 1959) for assessing convergent and discriminant validity has several limitations (cf. Bollen, 1989; Peter, 1981; Straub et al., 2004), other approaches have been suggested. Anderson and Gerbing (1988) state that discriminant validity is established when the value 1.0 is not in the confidence interval (± two standard deviations) around the correlation estimates of each pair of latent variables, or alternatively when the chi-square difference between two models – one which constrains the correlation between constructs to 1.0 (one pair of constructs at a time) and another which frees them – is significant, this indicates the two constructs are distinct. Another way to assess discriminant validity is to check whether within-construct correlations exceed between-construct correlations. If the two constructs are not highly correlated the above assumption would be expected to hold; however, if there is (expected) high correlation between the constructs, nothing can concluded from the difference between within- and between-construct correlations (Bollen & Lennox, 1991). In the case of a reflectively measured construct, within-dimension should exceed between-dimension correlations.

Besides assessing the whole construct of business performance, a test of the discriminant validity should also be conducted to assess whether the proposed dimensions of business performance are in fact distinct from each other. One can compare the model with more than one dimension with a model in which the correlation between two of the dimensions is restricted to 1, which is a special case of the general model (Hughes et al., 1986). A difference-of-χ^2 test would provide statistical evidence of whether the general model presents a significant improvement over the restricted model. If dimensions seem not to be distinct, then one might consider joining (two of) them together and follow the validation procedures discussed before. Or one might keep their individuality for clarity of the model, but the interpretation of the individual impacts should be conducted with care.

REPLICATION, GENERALIZABILITY AND CROSS-NATIONAL VALIDATION

In order to go from a particular experiment to general conclusions using inductive inference, some premises related to the degree of uncertainty should be observed. However, sufficient data is often not available in social sciences research or there is lack of information due to ignorance – thereby jeopardizing the probability theory because the required probability distributions cannot be accurately constructed.

In this case – which happens particularly if the sampling process does not comply with a probabilistic design –, the principles of epistemic uncertainty can be used, helping to reduce uncertainty with an increased state of knowledge or collection of more data. Although formal theories to handle uncertainties (evidence theory, possibility theory and interval analysis) have been proposed in the literature, it would be enough to check for external generalizability and stability as far as the business performance construct is concerned.

Therefore, the newly developed measure should additionally be assessed in order to show how generalizable and stable the relationships are across populations, research subjects, settings (countries, industries, types of firms etc.), and times (MacKenzie, 2003) and to understand the limits of the concept's applicability and usefulness.

Furthermore there should be an investigation as to whether success and failure should be considered two ends of the same scale or should be conceptualized as two discrete phenomena of a distinct nature (as speculated by Cavusgil & Zou, 1994; Matthyssens & Pauwels, 1996).

Investigation of cross-national equivalence is part of the quest for generalizability. It makes sense when it is expected that two countries are similar enough for the same dimensions of performance to be expected to be relevant and different enough for the possibility of substantial variation (Styles, 1998). Singh (1995) alerts that, before cross-national data collection is started, the researcher should address issues related to functional equivalence (does the focal construct serve the same function in different nations?), conceptual equivalence (is the construct expressed in similar attitudes or behaviors across nations?), and instrument equivalence (are the scale items, response categories and questionnaire stimuli interpreted identically across nations?). Then, after applying the necessary procedure of translation and back-translation of the questionnaire, the researcher should check for factorial similarity, factorial equivalence, and metric equivalence (Singh, 1995). Factorial similarity means that the same number of factors would show up in the two countries and the same items would be found to load on to the same factors. Factorial equivalence means that the respective factor loadings would be identical (or, better, it is not possible to reject the hypothesis that they are equal) across each national sample (note, however, that lack of strict factor equivalence (invariance) should be judged in terms of its practical or substantive importance, cf. Marsh & Hocevar, 1985). Metric equivalence means that both factor loadings and error variances are identical across national samples. Douglas and Craig (1983, as cited in Styles, 1998) alert that metric equivalence may not be reached because respondents may have different levels of familiarity and experience with the form or content of a given scale, which may lead to different scoring across populations or because there is a phenomenon of scalar inequivalence, whereby the score obtained from different populations differs in meaning and interpretation (due to cultural or other country-specific differences).

In the case of empirical research on business performance assessment, it is necessary to beware of survivor bias (Marsh & Swanson, 1984), that is, inadvertently failing to include those firms (or SBU's) that have ceased to exist and, as a consequence, not including respective relevant performance indicators, which might call generalizability into question.

Calculation of Aggregated Metrics

It may sometimes be necessary to devise a single aggregated metric that would summarize and represent the construct or its dimensions, for example, (i) to employ item-to-total correlations for the assessment of unidimensionality, (ii) to test construct validity (iii) to use it with some statistical techniques which do not accept latent variables but only manifest-like variables (e.g., regression analysis, analysis of variance (ANOVA), examination of correlations or examination of cross-tabulations, among others, cf. Kim & Mueller, 1978), or (iv) for ease of managerial interpretation where the construct may have to be categorized as **good** or **bad**, although the cut-off point between **success** and **failure** would still be arbitrarily defined by the researcher or the respondent (Styles, 1998).

However, a summary score is not necessary for some statistical analyses (e.g., S.E.M. or other techniques that accept latent variables plus an error term) or when all the researcher needs to know is the pattern of correlations between the latent construct and its antecedent and consequent variables. This is the case with many studies of business performance whose purpose is to identify factors that explain observed variance. All the researcher would need to know is the relative position of cases (is *A* better than *B*?) in order to judge the relative impact of explanatory factors or the temporal evolution (Churchill, 1979). It would not be necessary to provide meaning to the raw scores, expressed in whatever units.

There are no unequivocal or inherently correct **natural** measurement scales to attribute a **value** to many variables in the social sciences (Bollen, 1989), for example, business performance. Nonetheless, raw scores of the newly developed measure may need to be calculated and interpreted. This is usually done in terms of relative position (above or below the average, or at a given percentile). That is to say "meaning is imputed to a specific score in unfamiliar units by comparing it with the total distribution of the scores and this distribution is summarized by calculating a mean and standard deviation" (Churchill, 1979, p. 72). If, however, the researcher finds it necessary to calculate a final score, there are several possible aggregation techniques that elicit a single metric out of the various scores of the indicators such as: plain summation, unweighted average, weighted average, or factor scores. Careful judgment should be exercised before blindly employing these aggregation methods. If there are distinct strategic or operational implications of the **dimensions** or if they suffer distinct types or degrees of impact from the influencing factors or are affected by different factors, it would not make sense to aggregate them into a single raw score, because that would hide relevant information about the association of the construct with its nomological network of relationships. Moreover, there might exist some threshold (superior or inferior) so that some poor result in a given indicator (or dimension) may not be compensated by some good result in another (Ramaswamy, Kroeck, & Renforth, 1996).

Plain summation is calculated from a subset of items deemed to better represent the construct (or respective dimensions), usually those that load high on it. Such a score is usually not a perfect representation of the respective construct, be the nature of the indicators reflective or formative. In the case of reflective indicators, the linear composite of indicators differs from the respective construct not only in magnitude, but also because of measurement error (Bollen & Lennox, 1991). In the case of formative indicators, the linear composite would only equal the construct in the very unlikely case that all coefficients are equal to one and the measurement error is zero (Bollen & Lennox, 1991). As for unweighted average, the same problems reported for plain summation apply, but its use can be justified given that item-to-total correlation criteria (which is a test of internal consistency) calls for total score of the scale to be calculated by the unweighted sum of the item scores (Nunnally, 1978). Use of weighted average can solve the magnitude problem (both the construct and its linear composite would have the same scale, meaning that a one-unit change in one would lead to a one-unit change in the other) only if the weight is the same for all indicators and equal to the sum of the regression coefficients, but this does not eliminate the measurement error problem. Different individual weights can also be used, but the substantive reasoning to choose the respective weights should be explained. Such weights could correspond to the degree of importance of each indicator in the final scale, but how should this be judged? One possibility is to ask managers to distribute points across the items. Another is the degree of correspondence with a criterion variable, if such an appropriate criterion exists. In any case of summation (averaged or not), the use of standardized variables (instead of using their raw values) prevents variables with higher averages (on their respective original measurement scales) from weighing more on the resulting scale score than variables with lower averages (Ramaswamy et al., 1996).

Factor scores are a weighted average (based on factor loadings) of all variables (standardized or otherwise) loading on the factor, whereas a summated scale (be it plain summation or (weighted or unweighted) average) is based on only a pre-selected subset of variables, usually only those that load high on the respective factor (Hair et al., 1998). Kim and Mueller (1978) present several different paths to the estimation of factor scores, which will not be detailed here. They note that in practice the different methods lead to factor score estimates that are highly correlated.

FINAL CONSIDERATIONS

There seems to be a lack of continuous and complementary efforts to assess the quality of measurement instruments in business performance research, with different researchers coming up with their own conceptual framework (not always adequately justified) and their own operationalization. In fact, many pieces of research have used very simple definitions and procedures, which do not adequately capture the multifaceted nature of the phenomenon.

The validation of the measurement model should precede the testing of substantive relationships and different samples should ideally be used for the two phases (Venkatraman & Grant, 1986). In fact, Edwards and Bagozzi (2000) state that “[a] theory can be divided into two parts: one that specifies relationships between theoretical constructs and another that describes relationships between constructs and measures [observed scores]” (p. 155). A stronger link between the theoretical construct and its measurement lowers the chances of incorrectly rejecting a hypothesis due to “lack of correspondence between the measurements and the concepts that the measurements are intended to represent” (Bagozzi & Phillips, 1982).

This paper presented a well-substantiated set of procedures for developing and validating a new measurement model of the business performance construct. It is expected that the methodological discussion and the illustrative examples presented will motivate other researchers to engage in a joint research effort to develop a more substantively based, psychometrically sound and empirically validated model of the construct.

Although the guidelines presented in this paper were illustrated with the specific case of the business performance construct, they are general enough as to be applied to the development and validation of measures of other complex constructs in the social sciences.

Artigo recebido em 23.04.2007. Aprovado em 02.08.2007.

REFERENCES

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin*, *103*(3), 411-423.

Anderson, J., & Gerbing, D. (1991). Predicting the performance of measures in a confirmatory factor analysis with a pretest assessment of their substantive validities. *Journal of Applied Psychology*, *76*(5), 732-740.

Bagozzi, R. P., & Phillips, L. W. (1982). Representing and testing organizational theories: a holistic construal. *Administrative Science Quarterly, 27*, 459-489.

Barney, J. (1996). *Gaining and sustaining competitive advantage*. Reading, MA: Addison-Wesley Publishing Company.

Bass, B. M., Cascio, W. F., & O'Connor, E. J. (1974). Magnitude estimations of expressions of frequency and amount. *Journal of Applied Psychology, 59*(3), 313-320.

Bollen, K. (1989). *Structural equations with latent variables*. New York: John Wiley & Sons.

Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: a structural equation perspective. *Psychological Bulletin, 110*(2), 305-314.

Boyd, B. K., Gove, S., & Hitt, M. A. (2005). Construct measurement in strategic management research: illusion or reality? *Strategic Management Journal*, *26*(3), 239-257.

Briggs, S., & Cheek, J. (1986). The role of factor analysis in the evaluation of personality scales. *Journal of Personality, 54*(1), 106-148.

Cameron, K. (1986). Effectiveness as paradox: consensus and conflict in conceptions of organizational effectiveness. *Management Science*, *32*(5), 539-553.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81-105.

Carmines, E., & Zeller, R. (1979). Reliability and validity assessment. In J. Sullivan & R. Niemi (Eds.). *Sage university papers series on quantitative applications in the social sciences* (6th ed.). Beverly Hills, CA: Sage Publications.

Carneiro, J., Silva, J. da, Rocha, A. da, & Hemais, C. (2005). Conceptualisation and measurement of business performance: a multidimensional approach. *Proceedings of the Iberoamerican Academy of Management Annual Conference*, Lisbon, Portugal, 4.

Cavusgil, S. T., & Zou, S. (1994). Marketing strategy-performance relationship: an investigation of the empirical link in export market ventures. *Journal of Marketing, 58*(1), 1-21.

Churchill, G., Jr. (1979). A paradigm for developing better measures of marketing constructs. *Journal* *of Marketing Research*, *16*(1), 64-73.

Cox, E. III (1980). The optimal number of response alternatives for a scale: a review. *Journal of Marketing Research*, *17*(4), 407-422.

Dess, G. G., & Robinson, R. B, Jr. (1984). Measuring organizational performance in the absence of objective measures: the case of the privately held firm and conglomerate business unit. *Strategic Management Journal*, *5*(3), 265-273.

DeVellis, R. (1991). *Scale development: theory and applications*. Newsbury Park, CA: Sage Publications.

Diamantopoulos, A. (1999). Export performance measurement: reflective versus formative indicators. *International Marketing Review*, *16*(6), 444-457.

Diamantopoulos, A., & Siguaw, J. (2006). Formative versus reflective indicators in organizational measure development: A comparison and empirical illustration. *British Journal of Management*, *17*(4), 263-282.

Diamantopoulos, A., & Winklhofer, H. (2001). Index construction with formative indicators: an alternative to scale development. *Journal of Marketing Research*, *38*(2), 269-277.

Edwards, J., & Bagozzi, R. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods*, *5*(2), 155-174.

Epstein, S. (1983). Aggregation and beyond: some basic issues on the prediction of behavior. *Journal of Personality*, *51*(3), 360-392.

Gerbing, D. W., & Anderson, J. C. (1988). An updated paradigm for scale development incorporating unidimensionality and its assessment. *Journal of Marketing Research*, *25*(2), 186-192.

Fornell, C., & Bookstein, F. (1982). A comparative analysis of two structural equation models: LISREL and PLS applied to market data. In C. Fornell (Ed.). *A second generation of multivariate analysis* (Vol.1, pp. 289-324). New York: Praeger.

Hair, J., Jr., Anderson, R., Tatham, R., & Black, W. (1998). *Multivariate data analysis* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Hair, Jr., J., Black, W., Babin, B., Anderson, R., & Tatham, R. (2005). *Multivariate data analysis* (6th ed.). Upper Saddle River, NJ: Prentice Hall.

Hansen, G. S., & Wernerfelt, B. (1989). Determinants of firm performance: the relative importance of economic and organizational factors, *Strategic Management Journal*, *10*(5), 399-411.

Hinkin, T. (1998). A brief tutorial on the development of measures for use in survey questionnaires, *Organizational Research Methods*, *1*(1), 104-121.

Hughes, M., Price, R., & Marrs, D. (1986). Linking theory construction and theory testing: models with multiples indicators of latent variables. *Academy of Management Review*, *11*(1), 128-144.

Jarvis, C., Mackenzie, S., & Podsakoff, P. (2003). A critical review of construct indicators and measurement model misspecification in marketing and consumer research. *Journal of Consumer Research*, *30*(2), 199-218.

Jöreskog, K., & Goldberger, A. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, *10*(351), 631-639.

Katsikeas, C., Leonidou, L., & Morgan, N. (2000). Firm-level export performance assessment: review, evaluation and development. *Academy of Marketing Science, 28*(4), 493-511.

Kim, J.-O., & Mueller, C. W. (1978). Factor analysis: statistical methods and practical issues. In J. Sullivan (Ed.). *Quantitative applications in the social sciences* (Chapter 14). Beverly Hills, Calif.: Sage Publications.

Lissitz, R. W., & Green, S. B. (1975). Effect of the number of scale points on reliability: a Monte Carlo approach. *Journal of Applied Psychology, 60*(1), 10-13.

MacCallum, R., & Browne, M. (1993). The use of causal indicators in covariance structure models: some practical issues. *Psychological Bulletin*, *114*(3), 533-541.

MacKenzie, S. (2003). The dangers of poor construct conceptualization. *Journal of the Academy of Marketing Science*, *31*(3), 323-326.

MacKenzie, S. B., Podsakoff, P. M., & Fetter, R. (1991). Organizational citizenship behavior and objective productivity as determinants of managerial evaluations of salespersons' performance. *Organizational Behavior and Human Decision Processes*, *50*(1), 123-150.

March, J. G., & Sutton, R. I. (1997). Organizational performance as a dependent variable. *Organization Science*, *8*(6), 698-706.

Marsh, H., & Hocevar, D. (1985). Application of confirmatory factor analysis to the study of self-concept: first- and higher-order factor models and their invariance across groups. *Psychological Bulletin*, *97*(3), 562-582.

Marsh, T. A., & Swanson, D. S. (1984). Risk-return tradeoffs for strategic management/response. *Sloan Management Review*, *25*(3), 35-49.

Matthyssens, P., & Pauwels, P. (1996). Assessing export performance measurement. In S. T. Cavusgil & C. Axinn (Eds.). *Advances in International Marketing*, Greenwich, CT: JAI Press.

McGahan, A. M., & Porter, M. E. (1997). How much does industry matter, really? *Strategic Management Journal*, *18*(Special Summer Issue), 15-30.

Menezes, D., & Elbert, N. F. (1979). Alternative semantic scaling formats for measuring store image: an evaluation. *Journal of Marketing Research*, *16*(1), 80-87.

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

Peter, J. (1981). Construct validity: a review of basic issues and marketing practices. *Journal of Marketing Research*, *18*(2), 133-145.

Powell, T. C. (1996). How much does industry matter, an alternative empirical test. *Strategic Management Journal*, *17*(4), 323-334.

Ramaswamy, K., Kroeck, K. G., & Renforth, W. (1996). Measuring the degree of internationalization of a firm: a comment. *Journal of International Business Studies, 27*(1), 167-177.

Rindskopf, D., & Rose, T. (1988). Some theory and applications of confirmatory second-order factor analysis. *Multivariate Behavioral Research*, (23), 51-67.

Roquebert, J. A., Phillips, R. L., & Westfall, P. A. (1996). Market vs. management: what "drives" profitability? *Strategic Management Journal*, *17*(8), 653-664.

Rumelt, R. P. (1991). How much does industry matter? *Strategic Management Journal*, *12*(3), 167-85.

Schmalensee, R. (1985). Do markets differ much? *American Economic Review*, *75*(3), 341-350.

Schriesheim, C. A., Powers, K. J., Scandura, T. A., Gardiner, C. C., & Lankau, M. J. (1993). Improving construct measurement in management research: comments and a quantitative approach for assessing the theoretical content adequacy of paper-and-pencil survey-type instruments. *Journal of Management*, *19*(2), 385-417.

Schwab, D. (1980). Construct validity in organizational behavior. In B. Staw & L. Cummings (Eds.). *Research in Organizational Behavior* (Vol. 2, pp. 2-4). Greenwich, CT: JAI Press.

Singh, J. (1995). Measurement issues in cross-national research. *Journal of International Business Studies*, *26*(3), 597-620.

Smith, A. M. (1999). Some problems when adopting Churchill's paradigm for the development of service quality measurement scales. *Journal of Business Research*, *46*(2), 109-120.

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

Styles, C. (1998). Cross-cultural examination of export performance. *Journal of International Marketing*, *6*(3), 5-31.

Tehrani, M., & Noubari, R. (2005). A statistical conversion technique: objective and perceptive financial measures of the performance construct. *Organizational Research Methods*, *8*(2), 202-221.

Venkatraman, N. (1989). Strategic orientation of business enterprises: the construct dimensionality and measurement. *Management Science*, *35*(8), 942-962.

Venkatraman, N., & Grant, J. (1986). Construct measurement in organizational strategy research: a critique and proposal. *Academy of Management Review*, *11*(1), 71-87.

Venkatraman, N. & Ramanujam, V. (1986). Measurement of business performance in strategy research: a comparison of approaches. *Academy of Management Review*, *11*(4), 801-814.

Walker, B. A., Olson, J. C., Celsi, R. L. & Chow, S. (1992). Is construct validity a problem of measurement or theoretical generalization? A reply to Malhotra. *Journal of Business Research*, *25*(2), 187-195.

Wernerfelt, B., & Montgomery, C. A. (1988). Tobin's *q* and the importance of focus in firm performance. *American Economic Review*, *78*(1), 246-250.