Benchmarking by Item Response Theory (BIRTH)

A benchmarking method using IRT to build competitiveness scales for Brazilian technology higher education

Received 27 March 2013
Revised 3 November 2013
Accepted 3 December 2013

Juliano Anderson Pacheco and Dalton Francisco de Andrade
Universidade Federal de Santa Catarina, Florianopolis, Brazil, and
Antonio Cezar Bornia
Department of Production Engineering,
Universidade Federal de Santa Catarina, Florianópolis, Brazil

Abstract

Purpose – The purpose of this paper is to present a new method for benchmarking, which allows the construction of scales of competitiveness for the comparison of products using Item Response Theory (IRT).

Design/methodology/approach – Theoretically, the method combines classic benchmarking process steps with IRT steps and demonstrates through mathematical models how this technique can measure the competitiveness of products by means of a latent trait.

Findings – The IRT method uses the theories of psychometrics to measure the competitiveness of products through qualitative and quantitative interpretation of the tangible and intangible characteristics of those products. To demonstrate the application of the developed method, the items were constructed for teaching staff.

Research limitations/implications – The application of the developed method will increase the accuracy of assessments of the competitiveness of a product because this method uses a mathematical model of the IRT to evaluate the characteristics product that reflect market competitiveness. Items must be selected based on theories relevant to the product and/or expert opinion or customers.

Practical implications – The applicability of the method results in the construction of a scale in which items identify good practice with greater difficulty because they are represented in the same units that index competitiveness. Thus, managers of companies obtain knowledge about their products and the market, which allows them to assess their performance against their competitors and to make decisions regarding the continuous improvement of their production process and expansion of product characteristics.

Originality/value – This work presents a new method for benchmarking using a quantitative technique that enables measurement of the latent trait of "competitiveness" through robust mathematical models.

Keywords Competitiveness, Performance measurement, Benchmarking

Paper type Research paper

1. Introduction

The dynamics of the marketplace, which affect companies in the long run, is a result of forces that govern the relationships inherent in the production chain. These forces were modeled by Porter (1979) and are fivefold: the bargaining powers of suppliers and customers, the rivalry among existing competitors and the threats of new entrants and substitute products. In every field there are important considerations directly influencing the strategies of the companies operating in the market.

Benchmarking: An International Journal
Vol. 22 No. 5, 2015
pp. 945-962

1463-5771
DOI 10.1108/BIJ-03-2013-0035

In this organizational context, modeled by Porter, to remain competitive and meet the demands of the market, organizations must apply techniques for monitoring the stages of development of their products, considering the available inputs and desires of customers, and the level of competitiveness of their direct and indirect competitors. Such monitoring generates knowledge concerning the products and the market and assists managers in evaluating their performance against competitors, especially with respect to decisions for improving the production process and expanding product characteristics (Porter, 1985).

With the globalization of markets, the boundaries of performance and extended enterprises are increasingly facing competition in different business areas (agribusiness, industry, trade and services). The concept of competing business, according to Kotler and Keller (2006), is connected to the supply of products that meet the common needs of consumers, i.e., the existence of competition is intrinsically linked to solutions of a common market demand.

An end product can be defined as a material good or a tangible or intangible service that is produced by a company, able to meet a demand and attuned to the needs of the consumer market (Kotler and Keller, 2006). Companies looking to stand out in their market must develop differentiated features in their products that can be perceived by customers. The search for best practices, according to Camp (1989), consists of a coherent research process looking for new ideas, methods, practices and processes to adopt practices or adapt the good aspects and implement them to become the best of the best. This process is called benchmarking and was first used by Xerox Corporation in the late 1970s.

In this competitive environment, there is an intrinsic need for tools to identify the differential factors that companies use to compete or introduce their products into the marketplace. These tools allow companies to modify their products and remain competitive. Monitoring the evolution of products in a comparative way over time is critical for monitoring market best practices. Benchmarking is one such tool because it periodically identifies whether a company utilizes best market practices in its products and/or processes (Kotler and Keller, 2006).

Thus, the competitiveness of a product cannot be measured directly because there is not an appropriate instrument and a unit of measurement for this purpose; it is considered a latent trait (Wilson, 2005). The most appropriate method for measuring a latent trait is the Item Response Theory (IRT) (Embretson and Reise, 2000), through which a market practice is transformed into a measurable item. Thus, the IRT enables the construction of a measurement scale and is able to measure the competitiveness of products, making it possible to monitor competitiveness over time.

2. Research problem

Benchmarking is a process that generates learning from best practices to improve the competitiveness of a company. There are some variations in the definition of the benchmarking process (Camp, 1989; Spendolini, 1992; McNair and Leibfried, 1992; Watson, 1993; Garvin, 1993); however, issues raised by Watson (1993) define the principles of the benchmarking process well: What should we evaluate? Who do we evaluate? How do we perform the procedure? How do they perform the procedure?

However, all benchmarking process models in the literature (Ahmed and Rafiq, 1998; Fong *et al.*, 1998; Carpinetti and Melo, 2002; Anand and Kodali, 2008) originate from a single classic model developed and presented at Xerox by Camp (1989). For this work, we adopted the model of Fong *et al.* (1998), who developed a precursor to

the Xerox model. These authors proposed a systematic approach for improving performance to meet the needs and requirements of customers (Figure 1).

As shown in Figure 1, the benchmarking process begins by referring directly to the market to meet the needs of the customers, and the process is composed of ten steps that comprise five phases: planning (scope of and techniques for data collection), analysis (measurement of the differences and design of desired performance), integration (setting goals and communication of these goals across an enterprise), action (implementation of action plans and periodic reassessment) and maturity (integration of targets for the management of the company).

The benchmarking process is cyclical and is dependent upon feedback periodically. Benchmarking is defined according to the particular goals of companies, which are attuned to the wishes of customers, with respect to the continuous improvement of processes and the characteristics of their products.

In Anand and Kodali (2008), benchmarking applications are classified according to their origin, which can be academia, consulting groups or organizations. There is a prevalence of models developed by consultants, which shows the number of models that have been applied in enterprises. Regarding the analytical techniques used in benchmarking, according to Moriarty (2011), they primarily include data envelopment analysis (DEA), analytical hierarchy process (AHP), principal component analysis (PCA) and common factor analysis (CFA). In the same study, Moriarty developed

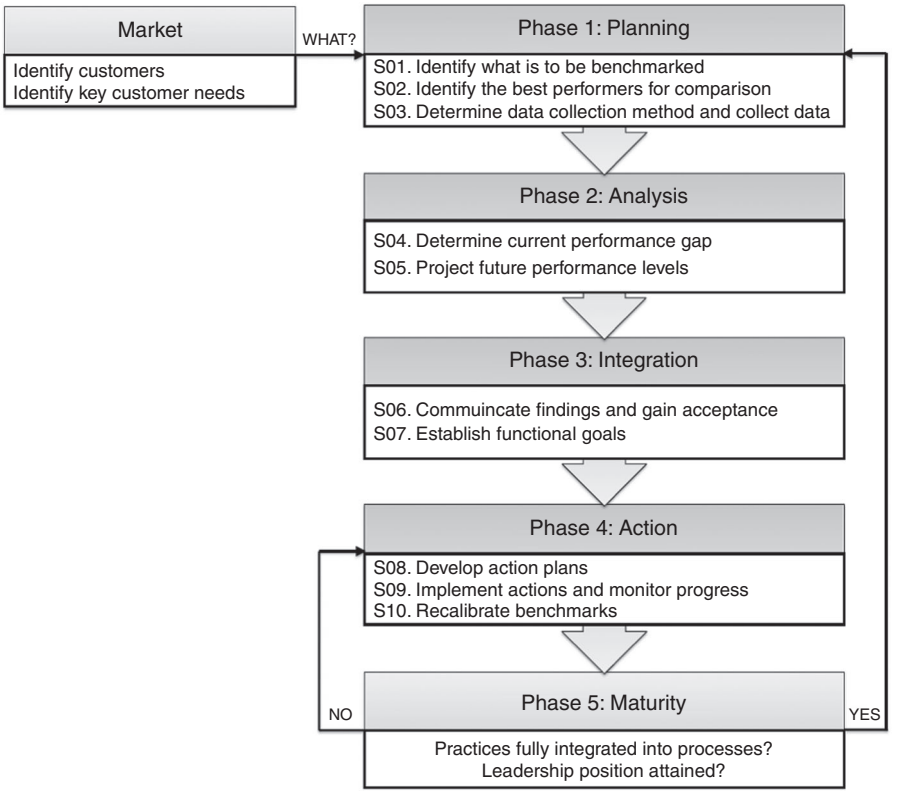

Figure 1. Benchmarking process reference

a theoretical framework that establishes the conditions necessary for benchmarking to be effective in business using axioms and logical conditions.

DEA is a multivariate approach, a contemporary process of benchmarking (Charnes *et al.*, 1978), data-oriented, which main objective is to evaluate the performance of a set of peer entities, called decision-making units (DMU). Examples of DMU: universities, banks, cities, states, countries, industries, hospitals, businesses in general (Cooper *et al.*, 2004). DEA has a major limitation, because it does not make possible to identify the individual contribution of the variables used in the analysis and does not allow tracking changes in efficiency over time (Habibov, 2010). Another difficulty is the lack of data in some cases, which must promote the elimination of some input variables in the analysis (Nayar, 2008). There is also a dependency of the results of the analysis of efficiency due to restrictions inherent to the problem (Hamdan, 2008).

AHP is a quantitative approach in the development of benchmarking as a support in the application of quality function deployment, since it allows complex decisions to be not based solely on experts' instinct. It also enables, in a quantitative way, defining the prioritization of features to be implemented, one of the main requirements in the benchmarking process (QFD Institute, 2013). To maximize the assertiveness of this technique, minimizing human uncertainty in assigning the weights, it is necessary to insert logic fussy (Büyüközkan, 2011, 2012;). To make the analysis more accurate, in assigning weights, one can use house of quality (Andronikidis, 2009). The multivariate techniques PCA and CFA are mainly focussed on dimensionality reduction, where the APC also seeks to find factors (Johnson and Wichern, 2002).

Thus, it is clear that benchmarking, besides the studies directed to its conceptual form focussing on understanding and changes in the application process (Fong *et al.*, 1998; Bhutta and Huq, 1999; Yasin, 2002; Dattakumar and Jagadeesh, 2003; Anand and Kodali, 2008; Moriarty, 2011), makes use of quantitative analysis techniques that have some limitations, generating more qualitative than quantitative results.

As a possibility to address the limitations identified in the technics DEA, AHP, PCA and CFA technics, we have the IRT, whose principles were developed in the 1920s (Bock, 1997), which aims at measuring a variable that cannot be observed directly, called latent trait (Wilson, 2005). IRT is a set of statistical models to measure a latent trace from the answers given to a questionnaire, consisting of a set of items related to the variable to be measured. This technique is widely used in psychological testing and educational assessments (Embretson and Reise, 2000), and more recently in areas such as life quality (Sijtsma *et al.*, 2008), psychiatry (Cúri *et al.*, 2011, and web usabilitity (Tezza *et al.*, 2011), for example.

In this context, this paper aims to introduce and develop a new method for benchmarking with the use of a quantitative technique, the IRT, which enables to build a scale of competitiveness and identify companies that, through their products, have the best market practices. This is done by creation and measuring of an index of competitiveness.

3. Methods

The procedure for creating a scale (Figure 2) to measure a latent trait was described by Embretson and Reise (2000): first, psychological system (theories concerning the latent trait); second, properties (attributes based on the theories of the latent trait); three, dimensional analysis (investigating factors); fourth, definition of the questionnaire (constitutive, based on other questionnaires, and operating with definitions to measure

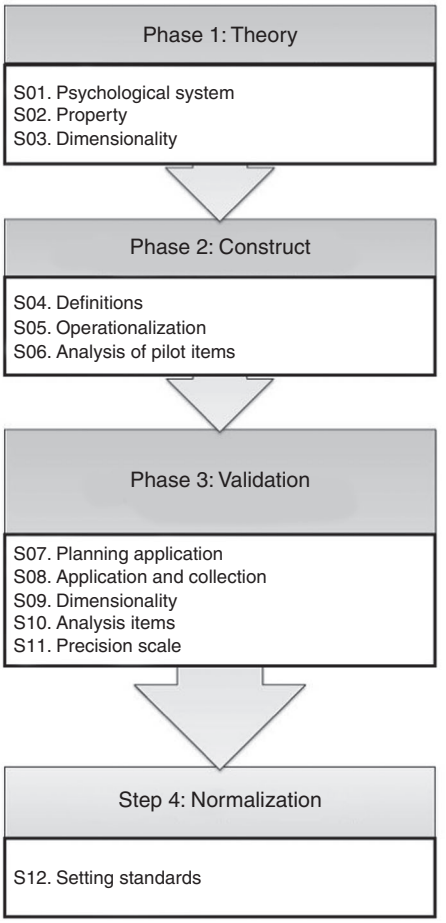

Figure 2. Process for developing a psychometric measure

the latent trait); fifth, operationalization of the questionnaire (construction of items from the relevant literature, interviews with experts, rules of construction, definition of criteria, and quantity of items); and six, theoretical analysis of the items (semantic analysis, understanding of items, and analysis of the judges).

This research proposes the inclusion of the procedure shown in Figure 2, the benchmarking process, while working with competitiveness as the latent trait.

3.1 Principles of IRT

In essence, IRT is a statistical technique that uses statistical models, with the aim of measuring a latent trait through the construction of a measuring scale that indirectly quantifies that trait. The construction of a scale makes quantification possible by selecting items that are suitable for measuring the latent trait. Upon completion of the analysis, the process provides feedback, which allows revision of the items and improvement of the questionnaire, thereby increasing the importance of translating the items that are valid for this measurement (Wilson, 2005).

A latent trait can be conceptualized as a hidden feature that cannot be measured directly but that can be observed in the responses to a questionnaire (Wilson, 2005). The two basic forms of modeling latent traits are the cumulative model of Guttman (1950) and the unfolding model of Coombs (1964).

The cumulative form relates primarily to the measurement of knowledge, as there is a direct relationship between the number of correct items and the knowledge of an individual, and this relationship determines the ability to accumulate items that model the latent trait.

Unfolding applies the measurement of attitudes, where the response of an item relates to a reaction, behavior or opinion of an individual with respect to certain objects, institutions, concepts or people. In this case, the latent trait represents if the individual is favorable or not those items.

The measurement of a latent trait is necessary to measure the items representing knowledge (cumulative) or attitude (unfolding). The latent trait is independent of the items and should not change when these items are updated because the new items can also measure the latent trait in question.

3.2 IRT models

There are four aspects that define which statistical model of the IRT should be used: first, the nature of the item; second, the number of people involved; third, the number of latent traits; and fourth, the characteristics of the cumulative or non-cumulative latent trait (Embretson and Reise, 2000).

The number of latent traits defines the dimensionality of the model and determines if the model uses a unidimensional or multidimensional trait. Each dimension represents a latent trait with its items and measuring scale. The items may be constructed to simultaneously measure two latent features, and this defines the multidimensionality of the items inside. When items measure only their latent traits, there is multidimensionality between items (Babcock, 2009).

However, there must be a line between the items and the latent trait, after the items represent the dimension/trait. These items can be divided into two groups: dichotomous and polytomous. The dichotomous items are binary and represent only two possibilities, such as right or wrong, present or absent, agree or disagree. Polytomous items must have at least three possibilities, and these can be nominal or ordinal. The difference between these two types is that the ordinal items possess an ordered relationship between the options and the nominal categories do not have a pre-established order.

Unidimensional models with dichotomous items are frequently used and in our application we will consider the two-parameter logistic model, usually referred as 2PL. This model is given by the following expression (Embretson and Reise, 2000):

$$P(U_{ij} = 1/\theta_j) = \frac{1}{1+e^{-a_i(\theta_j - b_i)}} \tag{1}$$

where: θ_j is the proficiency (latent trace) of respondent j; $P(U_{ij} = 1/\theta_j)$ is the probability of a correct response to item i, given the respondent's proficiency θ_j; a_i is the discrimination parameter of item i, with $a_i > 0$; b_i is the location parameter of item i, known as difficulty parameter, measured in the same metric(scale) as the latent trace.

The generalization of this model for two or more populations was developed by Bock *et al.* (1997). There are also unidimensional models for polytomous items.

The most popular is the Graded Response Model (Samejima, 1969), appropriate to model items with ordinal categories. The choice of the model depends on the latent trace one is modeling, unidimensional or multidimensional, and the items characteristics, dichotomous and/or polytomous. Details on multidimensional IRT models can be seen in Reckase (2009).

3.3 IRT process

The process of analysis involves two stages: a preliminary stage with Classical Test Theory (CTT) and the IRT analysis itself (Zimowski *et al.*, 2003).

CTT is the classical analysis of the items and is ensures the quality of the associated item with alternatives for specialists in the field and allows templates to identify incorrect or poorly formulated questions. It may be considered to be a pre-test before the IRT to identify potential problems with the items in the questionnaire. In CTT, the measure is the score, which uses characteristics that are cumulative and based on the number of correct answers. CTT does not allow the possibility of prediction. The item depends on who responds, and the number of scores defines the scale. These are the limitations of the classical theory. The following criteria are used in this analysis: the distribution of responses for the option and a high biserial coefficient (> 0.40), which can be confirmed by the IRT.

While the analysis by IRT aims to create a measurement scale for the latent trait, which cannot be directly measured, the challenge is the creation of a measuring instrument relevant to the latent trait, which is the questionnaire. The analysis is divided into two stages: the estimation of of the items parameters of the model, as presented in the previous section, and the estimation of proficiency, which is the quantification of the latent trait. The items parameters are independent from the respondents proficiency, and respondents' proficiencies are independent from the presented items and should not change with the exchange of other valid items. Ideally, the instrument must have items with high degrees of discrimination (parameter *a*) and with different difficulty settings (parameter *b*) to cover all the range of the scale (Baker and Kim, 2004).

After estimating the items parameters one can estimate the proficiency of each respondent and construct the scale by positioning the items on the proficiency scale. The construction of the scale is performed by the process called anchoring, and described in details in Beaton and Allen (1992). Given two successive anchoring levels x and y, with $x < y$, we position item i on level y if:

$$P(U_i = 1/\theta = y) \geqslant 0.65 \tag{2}$$

$$P(U_i = 1/\theta = x) \leqslant 0.50 \tag{3}$$

$$P(U_i = 1/\theta = y) - P(U_i = 1/\theta = x) \geqslant 0.30 \tag{4}$$

Using Expression (2), the probability of success, given proficiency y, is greater than 0.65 (65 percent). In Expression (3), the hit probability, given proficiency x, is less than or equal to 0.50 (50 percent), and the difference between these probabilities, as shown by (4), must be greater than 0.30 (30 percent). This means that there must be discrimination and group formation when there are items on the anchoring levels of the scale.

One can also compare the performance of different groups submitted to different instruments of measurement (questionnaires). All it is required is that the different questionnaires have some common items. This process is known as equating and details can be seen in Embretson and Reise (2000).

4. Findings

The goal of this new method is to evaluate the competitiveness of a product, by considering it a latent trait. This is the essence of the benchmarking method proposed here, in which the inability to directly measure competitiveness is resolved using an indirect measurement through IRT.

The principle for developing a new method is based on the benchmarking process of constructing psychometric scales (Figure 2). Using the principles of measurement of a latent trait, from psychometrics, the proposed new method is to apply the steps of the building process with IRT to psychometric scales, replacing the classic stages of benchmarking presented in Figure 1.

The integration of the theory, instrument analysis, and instrument validation and standardization form the entire process of applying IRT. The definition of the latent trait through dimensionality analysis, the construction of items, pilot study and application of the measuring instrument, along with an analysis of the items measuring the proficiency and scale construction comprises the new method. These steps will be cited simply as IRT, and the use of these steps replaces the steps of planning and benchmarking analysis. The remaining steps of the process (integration, action and maturity) will be implemented through the consolidated model in Figure 1. The proposed equivalence of the benchmarking steps (steps 1 through 5 – planning and analysis phases) with the steps of constructing a psychometric scale are presented in Figure 3.

The planning stage of this new method begins with the definition of the latent trait of the products from the companies that will be compared. The data collected follows the methodology for the items formulation. The result of the comparison will be presented by developing a scale that represents the level of competitiveness of the products considered.

4.1 Construction of items

The definition of items is one of the most important steps in the development of the benchmarking process with the use of IRT, and it shapes the competitiveness of products. In other words, the items represent the characteristics that represent a product's competitiveness in the market. The sources of the items may be the literature, interviews with experts or even customers themselves, as shown in the benchmarking process (Figure 1). To illustrate, the four basic types of items, shown in Table I, depend on the type of variable used to represent the features of a product.

Using the examples in Table I, it is possible to construct the same characteristic for different types of items. If characteristic X was the screen of an electronic product, this could be modeled by a binary item, such as whether the screen was colored or not; a nominal item, such as whether it was a Plasma, LCD or LED screen; ordinal item, such as the importance of having a color screen (low, medium or high); or an interval item (quantitative), such as the resolution of the screen in pixels.

Another point that should be considered in the definition of product items is the dimensionality of competitiveness because some products have more than one

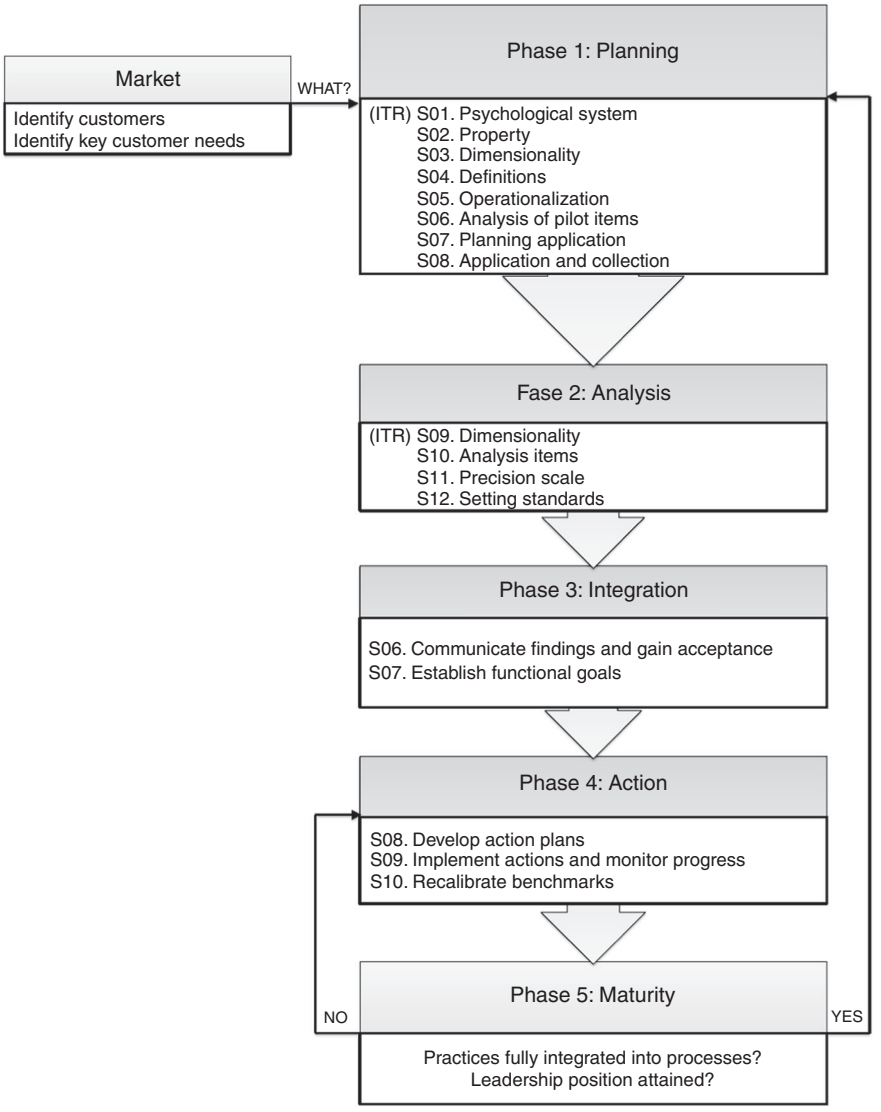

Figure 3.
New Benchmarking method by Item Response Theory (BIRTH)

Type of variable	Item type	Example item	Sample answers		
Binary	Dichotomic	It has the characteristic *X*?	Yes	No	
Nominal	Polytomous	What level of feature *X*?	Order 1	Order 2	Order 3
Ordinal	Polytomous	What is the name of the feature *X*?	Name 1	Name 2	Name 3
Interval	Polytomous	What is the value of feature *X*?	Value		

Table I.
Examples of items

dimension in their characteristics, which models their competitiveness. Basically, a product can have tangible characteristics (physical) with generally associated technical and technological features, and intangible characteristics, which usually related to services provided by the product.

Thus, one should consider the type of item (binary, nominal, ordinal or interval) connected to the definition of dimensionality, which will be considered to model the competitiveness of a product. Depending on these items, one should choose the IRT model that appropriately models competitiveness.

4.2 Model fitting

After the definition of the latent trait, the questionnaire and their respective items, the next step is to define the model or models more appropriate to measure the competitiveness of a product. Thus, as in educational assessment, cumulative models are perceived as the most suitable models for competitiveness. From the time a product enters the marketplace, "learn" and "respond" are the most difficult items to assess as the level of competitiveness increases over time. An illustration of how a mathematical model can represent the IRT latent trait of competitiveness is understood by the two-parameter logistic model presented in (1).

This model considers the companies' products as equivalent to the students in educational assessment, in which each product responds to a questionnaire consisting of items developed to measure the level of competitiveness. The interpretation of the parameters to model the competitiveness is as follows:

- U_{ij} is the existence of the feature (item) i for the product j: when $U_{ij} = 1$, the product j has the feature i; otherwise, $U_{ij} = 0$;
- θ_j is the competitiveness index inherent in the product j and represents the level of competitiveness of this product on the market;
- a_i represents the power of discrimination of the characteristic i, called the discrimination parameter; and
- b_i is the difficulty of having satisfied feature i, the difficulty parameter, and is on the same proficiency scale.

The final expression is the 2PL model, given by (1).

4.3 Construction of the scale

With the items parameters estimated in hand, one makes the construction of the scale and estimates the respondent's proficiency. Based on the theoretical premise that the difficulty parameters of the items use the same unit to measure proficiency, which represents the competitiveness of a product, the first stage is the identification of the anchor items. For the identification of the conditions (2), (3) and (4) must be satisfied; however, in many situations one cannot have too many items satisfying the three conditions and only two of them are usually considered in practice.

By anchoring the items, it is possible to interpret the estimated proficiencies of products, thereby creating subgroups and identifying the best market practices. For example, given the existence of a set of anchor items for the characteristics of products, the operating system can update this. Products may have more than one operating system (multi-platform) and may be able to update the software for the user. Considering all of the estimated proficiencies of a certain type of product, products that

cannot have their software updated by a user possess a proficiency below this set of anchor items. If the proficiency of a product is above this set of anchor items, then the product would have these characteristics. Generally speaking, the logic presented in the example is shown in Figure 4.

Similar to its use in educational assessments, wherein the value of proficiency identifying a student's knowledge by checking the items with difficulty values smaller than the proficiency of the learner, the proposed method can identify the best features of the product. This is possible with the construction of the scale; thus, the interpretation, based on the anchor items, identifies the characteristics that should be prioritized and implemented to increase the level of competitiveness (latent trait) of the product.

5. Experimental results

In Brazil, especially in the 2010s, there was a considerable increase in the market for higher education. In 2001, there were 1,378 higher education institutions (HEIs), and in 2010, there were 2,378 institutions. This represents an increase of over 70 percent. In terms of enrollment, the increase was over 110 percent, pushing the number of enrollments to 6,379,299 (Brasil, 2012a). During this period, the largest expansion in enrollment in higher education occurred in institutions that offered technology higher education courses (THECs). This increase indicates, to some extent, a movement toward greater investment in vocational and technological education, which had 69,797 enrollments in 2001. In 2010, enrollments reached 781,609, which represented an increase of over 1,000 percent (more than tenfold) (Brasil, 2012a). These numbers imply an increase in the competitiveness of these courses, and thus reinforce the importance of a method to determine the best practices of the market, which is the main purpose of this work.

5.1 The questionnaire

The questionnaire consists of items that represent the characteristics of these courses and their competitiveness in the market. The source of the items was the 2010 Census of Higher Education, developed by INEP/MEC and provided in microdata format (Brasil, 2012c). Considering only the THECs, with statistical processing microdata, the census collected data from 905 HEIs, 4,653 courses, 213,142 teachers and 1,067,462 students in this modality.

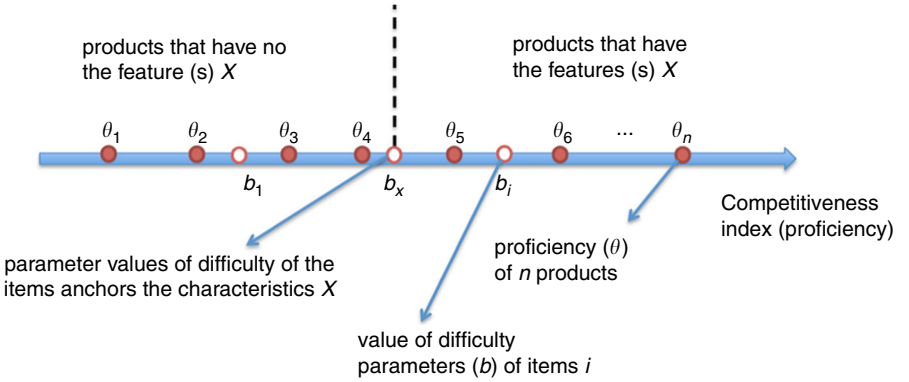

Figure 4. Interpretation of proficiency scale based on anchor items

A theoretical framework for the construction of the items was used as the basic current evaluation instrument for higher education courses in Brazil (Brasil, 2012b), which is part of the National System of Higher Education Assessment. In addition to being based on courses, the theoretical framework was also based on the evaluation of institutions and students (Brasil, 2009). The assessment instrument consists of three dimensions: organization didactic-pedagogical, teaching staff and infrastructure. With the data available from the census, 16 items were built. These items represent the teaching staff, as presented in Table II.

The submission of the courses, our respondents in this application, to the items of the questionnaire, as described in Table II, generated responses to the items, which were dichotomous and resulted in a database with binary variables. The settlement (assigned a score of "1") occurs when the answer is "yes" to the item; when the answer is "no" (assigned a score of "0").

To perform an analysis on a set of courses having similar characteristics within the same market segment, the data table was constructed using only classroom courses in science, mathematics, computing, engineering, production and construction, thus totaling 1,581 THECs.

Language R (R Development Core Team, 2012), a free software package from CTT (Willse and Shu, 2008) and Irtoys (Partchev, 2012), were used for the statistical analysis and the development of the classic and IRT steps. The software BILOG-MG (Zimowski *et al.*, 2003) was also used, specifically to estimate the parameters of the 2PL IRT model.

5.2 The competitiveness scale

With the definition of the latent trait competitiveness, as well as the questionnaire and its items, the most appropriate mathematical model to represent the competitiveness of THECs can be defined. The parameters of the mathematical models of the IRT represent the competitive in teaching staff dimension. Similar to the trend with the courses in the market, "learn" and "respond" are the most difficult items, as they increase their level of competitiveness over time; therefore, we used

CO_HEI		Identification HEI
CO_THEC		Identification THEC
HEI	ITEM01	The percentage of teachers spending > 40%?
Teachers	ITEM02	The teachers percentage with master's or doctorate degree > 50%?
	ITEM03	The teachers percentage with doctorate > 20%?
	ITEM04	The teachers percentage with full or partial dedication > 60%?
	ITEM05	The teachers percentage with full dedication > 20%?
	ITEM06	The teachers percentage who work in extension > 10%?
	ITEM07	The teachers percentage who work in management > 20%?
	ITEM08	The teachers percentage who work in graduate to distance > 2%?
	ITEM09	The teachers percentage who work in graduate classroom > 10%?
	ITEM10	The teachers percentage who work in research > 10%?
Students	ITEM11	The students percentage with internship > 50%?
	ITEM12	The students percentage with extension activities > 5%?
	ITEM13	The students percentage with monitoring activities > 5%?
	ITEM14	The students percentage with research activities > 5%?
	ITEM15	The percentage of locked students < 5%?
	ITEM16	The percentage of dropout students < 10%?

Table II.
Questionnaire for teaching staff dimension

a cumulative model. The mathematical model of the IRT most appropriate for the application is presented in (1). Establishing the relationship that THECs are equivalent students in educational assessment, the interpretation of the parameters are the same as presented in Section 4.2, where the product corresponds to educational services of the THEC.

The construction of the scale of competitiveness, followed by the analysis process of the IRT, is described in Section 4. In terms of the classical analysis, the criterion for checking the quality of the biserial coefficient is a high value (> 0.40), which is confirmed at the second stage by the IRT. In the analysis, when the Cronbach α was close to "1," the quality of the questionnaire is better. In the IRT, the goal is to estimate the parameters of the model presented in (1) and to estimate the proficiency, which consists of quantifying the latent trait, i.e., the competitiveness of the courses. In a quality questionnaire, considering model (1), the items must possess degrees of discrimination above 0.75 and different difficulty parameters; and in a range of competitiveness from −3 to +3.

In the analysis of the 16 items, nine showed the proper characteristics and were considered valid items by both the classical and IRT analyses; the summary of the analyses is shown in Tables III and IV. Cronbach's α, which measures the internal consistency of the items, is 0.72, a value that confirms the quality of the questionnaire.

			Classical analysis		
Items	Courses	Scores	% scores	Biserial coefficient	Cronbach's α deleted item
ITEM02	1,581	1,149	72.7	0.507	0.702
ITEM03	1,581	577	36.5	0.596	0.686
ITEM04	1,581	680	43.0	0.585	0.686
ITEM05	1,581	305	19.3	0.570	0.700
ITEM06	1,581	543	34.3	0.593	0.687
ITEM07	1,581	452	28.6	0.455	0.709
ITEM09	1,581	158	10.0	0.770	0.696
ITEM10	1,581	629	39.8	0.805	0.649
ITEM14	1,581	65	4.1	0.290	0.750

Table III. Classical analysis results

	IRT analysis			
	Discrimination (*a*)		Difficulty (*b*)	
Items	Estimation	Error	Estimation	Error
ITEM02	1.405	0.085	−0.944	0.057
ITEM03	1.540	0.087	0.510	0.043
ITEM04	1.385	0.080	0.275	0.044
ITEM05	1.288	0.088	1.436	0.082
ITEM06	1.788	0.098	0.553	0.039
ITEM07	1.224	0.078	0.958	0.064
ITEM09	2.495	0.173	1.571	0.057
ITEM10	5.026	0.285	0.271	0.018
ITEM14	0.798	0.122	4.310	0.572

Table IV. IRT analysis results

Following the estimation of the parameters, the next step was the construction of the scale, which entails identifying the anchors items, as shown in (2), (3) and (4). Figure 5 is a competitive scale with the anchor items identified.

Finally, the last step is to estimate the proficiency, which had its own competitiveness index, a value given by IRT, initially measured in the (0,1) scale, where 0 represents the mean and 1 the standard error of the proficiencies of the respondents. The index can be transformed to any numerical range. In the distribution of competitive indices of courses, the 2PL was estimated by the Expression (1) shown in Figure 6.

The value "0" is the average of the indices of competitiveness of the courses and any course with proficiency below that value is below the average of the analyzed courses. Based on Figure 5, 874 of the THECs have below average overall competitiveness indices, and the remaining 707 courses are above average. The interpretation of the competitiveness index of the courses is as follows (Figure 6):

- courses with index values below "0" are below average and do not possess any characteristics described by the anchor items; these conditions were found in 874 courses;
- courses with content in the range [0, +1] must have teachers with a master's or doctorate degree (ITEM02), and these conditions were present in 509 courses;

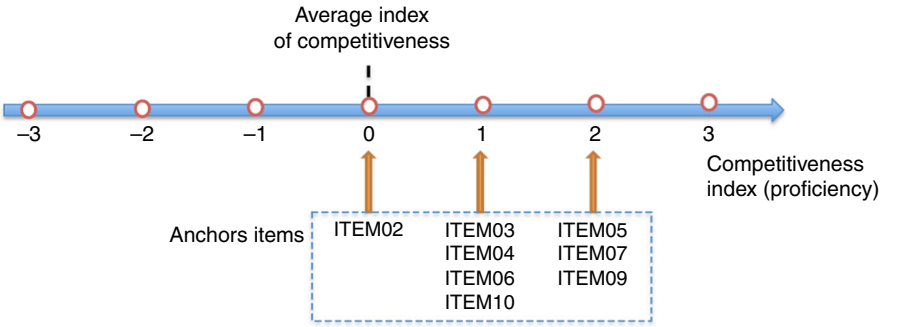

Figure 5.
Competitiveness scale of THECs and anchors

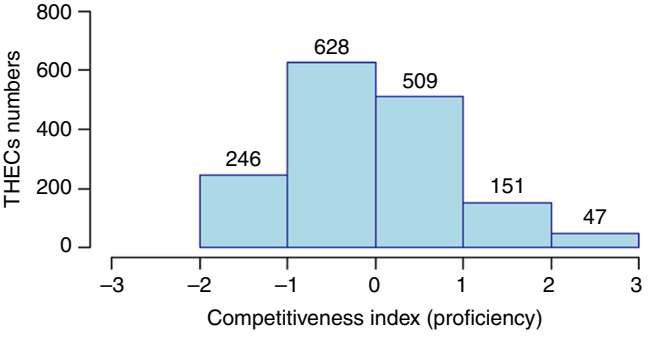

Figure 6.
Distribution of competitiveness index estimated by the model ML2

- courses with content on the interval [+1, +2] differ from the previous interval by having teachers with a doctorate (ITEM03) and dedication full/partial (ITEM04) with operations in extension (ITEM06) and research (ITEM10); this included a total of 151 courses; and
- courses with a competitiveness index above +2 differ from the others by having teachers with exclusive dedication (ITEM05) and operating in management (ITEM07) and graduate classroom (ITEM09), found in 47 courses.

One can see that there were results relevant to the area of knowledge, but we would like to emphasize that this application is just one example of the use of the method BIRTH, which in this case is evaluating the competitiveness of a product of type service. The method can also be applied to physical products and companies in general.

6. Conclusions

A new benchmarking process based on IRT is presented in this work, which meets the technical limitations of the two main techniques DEA and AHP usually applied in benchmarking process, without losing the advantages of these. Its relevance to the development of research and assessment of product competitiveness was shown through an application to a real data situation. In this new process, the interpretation of the items on the constructed scale allows managers to first identify the importance of a particular characteristic (item) and then the difficulty of insertion of a good practice (item) in the product. With the use of the IRT+benchmarking method, the manager can then determine which good practice should be used for their product to increase its competitiveness.

References

Ahmed, P.K. and Rafiq, M. (1998), "Integrated benchmarking: a holistic examination of select techniques for benchmarking analysis", *Benchmarking for Quality Management and Technology*, Vol. 5 No. 3, pp. 225-242.

Anand, G. and Kodali, R. (2008), "Benchmarking the benchmarking models, *Benchmarking: An International Journal*, Vol. 15 No. 3, pp. 257-291.

Andronikidis, A. (2009), "The application of quality function deployment in service quality management", *The TQM Journal*, Vol. 21 No. 4, pp. 319-333.

Babcock, B.G.E. (2009), "Estimating a noncompensatory IRT model using a modified metropolis algorithm", PhD dissertation, University of Minnesota, Minneapolis, MN.

Baker, F.B. and Kim, S-H. (2004), *Item Response Theory: Parameter Estimation Techniques*, Marcel Dekker, New York, NY.

Beaton, A.E. and Allen, N.L. (1992), "Interpreting scales through scale anchoring", *Journal of Educational Statistics*, Vol. 17 No. 2, Special Issue: (National Assessment of Education Progress), pp. 191-204.

Bhutta, K.S. and Huq, F. (1999), "Benchmarking – best practices: an integrated approach", *Benchmarking: An International Journal*, Vol. 6 No. 3, pp. 254-268.

Bock, R.D. (1997), "A brief history of item theory response", *Educational Measurement: Issues and Practice*, Vol. 16 No. 4, pp. 21-33.

Bock, R.D., Zimowski, M.F., Van Der Linden, W.J. and Hambleton, R.K. (1997), *Multiple Group IRT in Handbook of Modern Item Response Theory*, Spring-Verlag, New York, NY.

Brasil (2009), "Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP)", SINAES – Sistema Nacional de Avaliação da Educação Superior: Da Concepção à Regulamentação, Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, Brasília, 5ª ed, Disponível em: www.publicacoes.inep.gov.br/detalhes.asp?pub=4389 (acesso em: 20 julho 2011) (in Portuguese).

Brasil (2012a), "Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP)", Censo da educação superior: 2010 – resumo técnico, Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, Brasília, Disponível em: http://portal.inep.gov.br/web/censo-da-educacao-superior/resumos-tecnicos (acesso em 20 julho 2011) (in Portuguese).

Brasil (2012b), "Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP)", Instrumento de Avaliação de Cursos de Graduação presencial e a distância, Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, Brasília, Disponível em: http://portal.inep.gov.br/superior-condicoesdeensino-manuais (acesso em: 20 julho 2011) (in Portuguese).

Brasil (2012c), "Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP)", Microdados para Download, Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, Brasília, Disponível em: http://portal.inep.gov.br/basica-levantamentos-acessar (acesso em 20 julho 2011) (in Portuguese).

Büyüközkan, G. (2011), "Strategic analysis of healthcare service quality using fuzzy AHP methodology", *Expert Systems with Applications*, Vol. 38 No. 8, pp. 9407-9424.

Büyüközkan, G. (2012), "A combined fuzzy AHP and fuzzy TOPSIS based strategic analysis of electronic service quality in healthcare industry", *Expert Systems with Applications*, Vol. 39 No. 3, pp. 2341-2354.

Camp, R.C. (1989), *Benchmarking: The Search for Industry Best Practices that Lead to Superior Performance*, ASQC Quality, Milwaukee, WI.

Carpinetti, L.C.R. and Melo, A.M. (2002), "What to benchmark?: A systematic approach and cases", *Benchmarking: An International Journal*, Vol. 9 No. 3, pp. 244-255.

Charnes, A., Cooper, W.W. and Rhodes, E. (1978), "Measuring the efficiency of decision making units", *European Journal of Operational Research*, Vol. 2 No. 6, pp. 429-444.

Coombs, C.H. (1964), *A Theory of Data*, Wiley, New York, NY.

Cooper, W.W., Seiford, L.M. and Zhus, J. (2004), *Handbook on Data Envelopment Analysis*, Kluwer Academic Publisher, Boston, MA.

Cúri, M., Singer, J.M. and Andrade, D.F. (2011), "A model for psychiatric questionnaires with embarrassing items", *Statistical Methods in Medical Research*, Vol. 20 No. 5, pp. 451-470.

Dattakumar, R. and Jagadeesh, R. (2003), "A review of literature on benchmarking", *Benchmarking: An International Journal*, Vol. 10 No. 3, pp. 176-209.

Embretson, S. and Reise, S.P. (2000), *Item Response Theory for Psychologists*, Lawrence Erlbaum Associates Inc. Publishers, Mahwah, NJ.

Fong, S.W., Cheng, E.W.L. and Ho, D.C.K. (1998), "Benchmarking: a general reading for management practitioners", *Management Decision*, Vol. 36 No. 6, pp. 407-418.

Garvin, D.A. (1993), "Building a learning organization", *Harvard Business Review*, Vol. 71 No. 4, pp. 78-91.

Guttman, L. (1950), "The basis for scalogram analysis", in Stouffer, S.A., Guttman, L., Suchman, E.A., Lazarsfeld, P.F., Star, S.A. and Clausen, J.A. (Eds), *Measurement and Prediction, Studies in Social Psychology in World War II*, Vol. 4, Princeton University Press, Princeton, NJ, pp. 60-90.

Habibov, N.N. (2010), "Comparing and contrasting poverty reduction performance of social welfare programs across jurisdictions in Canada using data envelopment analysis (DEA):

an exploratory study of the era of devolution", *Evaluation and Program Planning*, Vol. 33 No. 4, pp. 457-467.

Hamdan, A. (2008), "Evaluating the efficiency of 3PL logistics operations", *Int. J. Production Economics*, Vol. 113 No. 1, pp. 235-244.

Johnson, R.A. and Wichern, D.W. (2002), *Applied Multivariate Statistical Analysis*, Prentice-Hall, Upper Saddle River, NJ.

Kotler, P. and Keller, K.L. (2006), *Marketing Management*, 12th ed., Prentice Hall, Upper Saddle River, NJ.

McNair, C.J. and Leibfried, K.H.J. (1992), *Benchmarking, A Tool for Continuous Improvement*, Harper Business Press, New York, NY.

Moriarty, J.P. (2011), "A theory of benchmarking", *Benchmarking: An International Journal*, Vol. 18 No. 4, pp. 588-612.

Nayar, P. (2008), "Data envelopment analysis comparison of hospital efficiency and quality", *J Med Syst*, Vol. 32 No. 3, pp. 193-199.

Partchev, I. (2012), "Irtoys: simple interface to the estimation and plotting of IRT models", R package version 0.1.6, available at: http://cran.r-project.org/package.irtoys (acessed 2 January 2013).

Porter, M.E. (1979), "How competitive forces shape strategy", *Harvard Business Review*, March/April, pp. 137-145.

Porter, M.E. (1985), *Competitive Advantage*, Free Press, New York, NY.

QFD Institute (2013), "The official source for QFD", available at: www.qfdi.org/index.html (acessed 2 February 2013).

R Development Core Team (2012), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, available at: www.r-project.org (acessed 2 January 2013).

Reckase, M.D. (2009), *Multidimensional Item Response Theory*, Springer, New York, NY.

Samejima, F. (1969), "Estimation of latent ability using a response pattern of graded scores", *Psychometrika Monograph Supplement*, Vol. 1 No. 17, pp. 1-100.

Sijtsma, K., Emons, W.H.M., Bouwmeester, S., Nyklícek, I. and Roorda, L.D. (2008), "Nonparametric IRT analysis of Quality-of-Life Scales and its application to the World Health Organization Quality-of-Life Scale (WHOQOL-Bref)", *Qual Life Res.*, Vol. 17 No. 2, pp. 275-290.

Spendolini, M.J. (1992), *The Benchmarking Book*, AMACOM, New York, NY.

Tezza, R., Bornia, A.C. and Andrade, D.F. (2011), "Measuring web usability using item response theory: principles, features and opportunities", *Interacting with Computers*, Vol. 23 No. 2, pp. 167-175.

Watson, G.H. (1993), *Strategic Benchmarking: How to Rate your Company's Performance Against the World's Best*, John Wiley & Sons Inc., New York, NY.

Willse, J.T. and Shu, Z. (2008), "CTT: Classical Test Theory Functions", R package version 1.0, available at: http://cran.r-project.org/package.ctt (acessed 2 January 2013).

Wilson, M. (2005), *Constructing Measures: An Item Response Modeling Approach*, LEA, London.

Yasin, M.M. (2002), "The theory and practice of benchmarking: then and now", *Benchmarking: An International Journal*, Vol. 9 No. 4, pp. 217-243.

Zimowski, M.F., Muraki, E., Mislevy, R.J. and Bock, R.D. (2003), *BILOG-MG 3 for Windows: Multiple-group IRT Analysis and Test Maintenance for Binary Items [Computer Software]*, Scientific Software International Inc., Skokie, IL, available at: www.ssicentral.com (acessed 2 January 2013).

Further reading

Donnelly, M. (2000), "A radical scoring system for the european foundation for quality management business excellence model", *Managerial Auditing Journal*, Vol. 15 Nos 1/2, pp. 8-11.

Grupp, H. (2004), "Indicators for national science and technology policy: how robust are composite indicators?", *Research Policy*, Vol. 33 No. 9, pp. 1373-1384.

Ho, W. (2006), "Multiple criteria decision-making techniques in higher education", *International Journal of Educational Management*, Vol. 20 No. 5, pp. 319-337.

Leong, W.H. (2003), "Measuring the technical efficiency of banks in Singapore for the period 1993-99: an application and extension of the Bauer *et al.* (1997) technique", *ASEAN Economic Bulletin,* Vol. 20 No. 3, pp. 195-210.

Mehregan, M.R. (2010), "An optimisational model of benchmarking", *Benchmarking: An International Journal*, Vol. 17 No. 6, pp. 876-888.

Naor, M. (2010), "The globalization of operations in Eastern and Western countries: Unpacking the relationship between national and organizational culture and its impact on manufacturing performance", *Journal of Operations Management*, Vol. 28 No. 3, pp. 194-205.

Pasiouras, F. (2010), "Multicriteria classification models for the identification of targets and acquirers in the Asian banking sector", *European Journal of Operational Research*, Vol. 204 No. 2, pp. 328-335.

Potnis, D.D. (2010), "Measuring e-Governance as an innovation in the public sector", *Government Information Quarterly*, Vol. 27 No. 1, pp. 41-48.

Tan, H.B. (2008), "Relative efficiency measures for the knowledge economies in the Asia Pacific region", *Journal of Modelling in Management*, Vol. 3 No. 2, pp. 111-124.

Corresponding author
Professor Juliano Anderson Pacheco can be contacted at: jap@deps.ufsc.br