REMark – Revista Brasileira de Marketing
e-ISSN: 2177-5184
DOI: 10.5585/remark.v13i2.2685
Data de recebimento: 08/01/2014
Data de Aceite: 17/03/2014
Editor Científico: Otávio Bandeira De Lamônica Freire
Avaliação: Double Blind Review pelo SEER/OJS
Revisão: Gramatical, normativa e de formatação

MEASUREMENT IN MARKETING: CURRENT SCENARIO, RECOMMENDATIONS AND CHALLENGES

ABSTRACT

The purpose of this article is to discuss about construct measurement in Marketing by summarizing the main considerations about the subject. First, it discusses the origins of the debates about the theme since the 1970s and describes its main consolidated models (the classical Churchill's model, the COARSE model and the formative measurement model). Then it presents current concerns about the classical approach with relevant recommendations (particularly regarding multi-item measurement, single-item measurement, rating scales and cross-cultural aspects). At the end, it presents considerations about measurement trends in Marketing with emphasis on the Item Response Theory (IRT), Bayesian estimators and Partial Least Squares (PLS). The article updates the debate on the theme and contributes to Marketing experts and researchers who demand a current view about measurement and recommendations for research development.

Keywords: measurement in marketing, constructs, marketing scales, validity, reliability

MENSURAÇÃO EM MARKETING: ESTADO ATUAL, RECOMENDAÇÕES E DESAFIOS

RESUMO

Este artigo tem por finalidade debater o tema de mensuração de construtos em Marketing, sumarizando as principais discussões sobre o assunto. Inicialmente, discutimos a origem das preocupações e os desdobramentos na área desde os anos de 1970. Em seguida, apresentamos os principais modelos consolidados (modelo clássico de Churchill, modelo COARSE e modelo de mensuração formativa). Na sequência, apresentamos preocupações atuais que se somam à teorização clássica, com algumas recomendações relevantes (especialmente sobre mensuração por múltiplos itens, mensuração por um único item, escalas de verificação e aspectos transculturais). Ao final, apresentamos considerações sobre tendências de mensuração em Marketing, com ênfase em Teoria da Resposta ao Item, operadores Bayesianos e estimação por mínimos quadrados parciais. O artigo atualiza o debate sobre o tema e tem a possibilidade de contribuir para estudiosos e pesquisadores de Marketing que demandem uma visão atual sobre mensuração e recomendações para pesquisas.

Palavra chave: Mensuração; Escalas; Validação; Confiabilidade

Felipe Zambaldi[1]
Francisco José da Costa[2]
Mateus Canniatti Ponchio[3]

[1] Doutor em Administração de Empresas pela Fundação Getulio Vargas - FGV. Professor da Fundação Getulio Vargas – FGV, Brasil. E-mail: **felipe.zambaldi@fgv.br**

[2] Doutor em Administração de Empresas pela Fundação Getulio Vargas – FGV. Professor da Universidade Federal da Paraíba, UFPB, Brasil. E-mail: **franzecosta@gmail.com**

[3] Doutor em Administração de Empresas pela Fundação Getulio Vargas – FGV. Professor da Escola Superior de Propaganda e Marketing de São Paulo (ESPM-SP), Brasil. E-mail: **mponchio@espm.br**

1 INTRODUCTION

The scientific knowledge construction process is largely dependent on the researchers' ability to properly measure the concepts they address. Unlike some scientific areas where most concepts can be directly observed (such as height, weight and age), we find in social sciences, and particularly in Marketing, constructs of an abstract nature that cannot be directly accessed, like satisfaction, loyalty, happiness, materialism and brand attitude.

Measuring values, beliefs and attitudes depends, at first, on a great effort of concept definition (for instance, what exactly do we mean when we mention satisfaction?). And after eliminating the communication obstacle of concept clarity, we need a measurement strategy. Our objective is to place units of analysis (products, consumers and companies, for instance) on one axis according to their level of a certain characteristic of interest being measured, that is, we need to define a system to measure intensity (or quantity) of the construct we have defined.

For example, how do we measure intelligence? Even if we have a consensual definition about what intelligence is (a quick search in the literature will bring complementary visions about the construct), there is no tag on people's bodies indicating their degree of intelligence. This concept of a latent nature (it is present in the object, but we do not see it) cannot be directly measured and, therefore, should be accessed by means of indirect measurement strategies.

In this article, we attempted to produce a *tour de force* about measurement practices for Marketing, a field that is typically interested in assigning values to concepts that are not directly observable, for subsequent statistical operationalization of data, which we generate to analyze assumptions involving the constructs. For this purpose, we first place the problem of measuring abstract and latent constructs in both historical and current perspectives, by presenting the classical approaches, and more recent developments, in order to introduce a contemporaneous debate on the theme. After that, we provide procedures and recommendations for scales development, for the evaluation of instrument validity and reliability, and to address rating scales (coherence between scale and content; number of points; aggregation strategy; and use of statistical techniques). We focus on the provision of alternatives for formative indicators, as well as reflective indicators, more commonly found in the literature. In particular, we attempt to provide readers with content to keep a discussion about aspects for the elaboration and use of scales in cross-cultural studies, pointing out required adaptations to scales when applying them in distinct contexts, and discuss about measurement trends in Marketing based on current debates and their responses to the fragilities of more usual models. We address specifically the Item Response Theory (IRT), Bayesian estimators and the models of Partial Least Squares (PLS).

2 THE HISTORICAL AND CURRENT PROBLEM OF ABSTRACT AND LATENT CONSTRUCT MEASUREMENT

In an interesting article with an overview of Statistics, Pereira (1997) highlights that measurement is one of the core elements of the statistical process (called by the author as the 'technology of science'). For Pereira, the conventional scientific process, which develops the empirical evaluation of propositions and hypotheses, successively goes through the decision to measure variables of interest in the empirical field, data collection for the measurement of scales, and across the analysis of such data, a stage which applies several statistical techniques.

The research structure reported above - also considered by other authors (see Pedhazur & Schmelkin, 1991) - shows that researchers need to consider these procedures (measurement, design and analysis) as a reference for knowledge construction. It seems that the general emphasis of research on social and behavioral sciences has historically favored the analysis dimension, with greater rigor on statistical analysis techniques.

Starting in the 1960s, the measurement analysis as a core element of the quantitative research process in Marketing reached a different status worldwide. In Brazil, this tendency consolidated more recently, starting in 2000, as a natural evolution of more academic research that Business Administration schools have adopted since. The measurement analysis is today requested in most research reports presented as dissertations, theses and articles.

Actually, the Marketing research field has absorbed a longtime recurring concern in the fields of Education and Psychology, contexts in which measurement has been an object of study and analysis for more than one century. The reason for such absorption is simple: we work in Marketing with abstract constructs (such as satisfaction, identity, attachment, loyalty), which we assume to be measurable, but for which we still have no instruments to access directly. That happens in the assumption that a measurable degree exists for stress (in Psychology) or knowledge (in Education), for instance, but we do not have any instrument to directly access these constructs. That is, we study latent constructs in Marketing that require their own and different measurement strategy like those used, for example, in the area of Finances to measure profit, or in the area of Production to measure quality problems.

We have absorbed in Marketing most of the substantive content of the measurement theory used in Psychology and Education, to enable, more recently, a more adequate contribution of the content we produce. To build a basic reference of what is already consolidated in Marketing, we briefly bring some information about these two fields.

In Psychology, the measurement problem started when professionals of this area decided to develop tests (metrics) to evaluate constructs and variables. The psychological testing field (see Urbina, 2004) and the discipline of Psychometrics have attempted to develop tests and methods since the end of the 19th century to measure, for instance, personal values, professional trends or predisposition to certain behaviors, with instruments of pencil and paper (or equivalent, such as today's scanned instruments applied via internet). In this field, one of the main academic journals on measurement, called *Psychometrika*, created in 1936, has provided several theoretical contributions that go beyond the field of Psychology.

In Education, the measurement problem affects most people attending school, once the school tests are tentative instruments of learning measurement applied by teachers during their classes. In this field, measurement is reported as a core element of the specialized area of Educational Evaluation, which includes both learning evaluation for the knowledge transferred by teachers and the evaluation of competences (such as in public contests) and the evaluation of programs and institutions (like institutional evaluations and postgraduate program evaluations). It was in the field of Education that the most recent and relevant studies about the Item Response Theory (IRT) have been developed, which we discuss later.

Although no reference date has been defined, we can say that, in Marketing, the first significant step for the definition of a measurement priority is in the article of Gilbert Churchill, published in 1979 by the reputable *Journal of Marketing Research*, which brought a well-grounded criticism of practices in Marketing research adopted those days, which, according to the author, were extremely fragile. The warning was simple at that time, but it is still valid: it is not possible to believe in the value of numbers operationalization if we do not know for sure what is behind these numbers (that is, in the decisions regarding measurement and design for data collection).

Churchill rescued the previous constructions in the academic studies in Marketing[4], Psychology and Education, and proposed a procedure to be used by researchers when creating metrics. His model has been recurrently mentioned and used in Marketing researches (when this article was elaborated, there were more than 9600 citations of his articles in Google Scholar), but it had limitations and received criticisms.

Churchill's model is focused on the development of measures based on some assumptions that, if sometimes not valid, can lead to the proposition of other models of metric constructions. His main recommendations were regarding the following: in the principle of measurement, according to the domain sampling theory, multiple indicators are always used to measure a construct, and the validation analysis can be conducted by using techniques such as factor analysis (for the identification or reaffirmation of an underlying factor – the latent construct – explaining the variation of items), and the Cronbach's coefficient alpha (to attest the internal consistency in the set of items). When Churchill's principles (or their general application) were denied, other developments proposed single-item measurement or/and a qualitative analysis as validation (especially defended by John Rossiter in his COARSE model), or formative measurement, which does not assume an underlying factor explaining the variation of a group of indicators, but inversely, assumes that the variation of items implies the variation of the construct (several studies defend this controversial thesis, such as the text of Diamantopoulos and Winklhofer, 2011).

By means of the analysis of recent studies and publications, we can say that the current scenario focuses on the discussion of these three perspectives: classical model (based on Churchill's model, 1979); formative measurement; and measurement with less quantitative elaboration, with a greater focus on qualitative validation. Other developments seem to be the subject of future studies and applications, like the expansion of the use of the Item Response Theory, which widely applies to the field of educational evaluation and gradually enters the Marketing universe (see Andrade, Tavares & Valle, 2000; Lucian, 2012).

3 THE PROBLEM OF SCALE CONSTRUCTION: CLASSICAL ALTERNATIVES AND DEVELOPMENTS

To illustrate particularities in knowledge construction in different scientific fields, Mari (2005) compared the use of axioms in formal science (for example, the Euclidean geometry in which the elements of theory construction are founded on axioms) to the reliance on the empirical measurement of scientific phenomena. The author says that, in empirical sciences, different epistemological understandings coexist among scientists in terms of measurement and even of the possibility to assign a number to a phenomenon.

[4] In 1965, Charles Lee discussed the issues of measurement in a broader context of quantitative research and its difficulties and specificities; cf. Lee (1965).

3.1. The classical Churchill's model

Particularly in Marketing studies, the proposal of Gilbert Churchill and its derivations have predominated since 1979, creating the so-called classical approach of Marketing measurement. Churchill proposed these procedures after he realized that measurement efforts lacked rigor. In this context, the author presented definitions of validity and reliability, indubitably the two most important definitions for the measurement instrument validation process. These definitions of validity and reliability provided by Churchill are still adopted by most researchers in Marketing.

The author defines validity as the ability of a measurement to capture in its scores the phenomenon that it purports to measure, and reliability as the property of consistent measurements of the same construct. That is, validity ensures that the scale measures what it is supposed to measure, and reliability ensures that this measurement with minimum random errors (which are expected in the scientific process, but need to be minimized).

Churchill's proposal for measurement validation consists of sequential steps, some of which may repeat in the same process. The first step involves the specification of the theoretical construct domain, or its theoretical definition, and it should be based on literature analysis. The next step refers to the generation of a set of items (questions) that will constitute the first version of the measurement instrument. This stage depends on the previous stage (construct domain specification) and is based on literature analyses, evaluation of empirical studies published previously, creation of examples and incidents that are relevant to the conceptual domain, and qualitative studies with target raters, by means of focus groups, for instance. After generating the first set of items, data collection is conducted for a pre-test. Based on the pre-test results, the next stage will refine the instrument to check which items should remain and which items should be excluded or adapted. The tools proposed by Churchill for this stage are: the Cronbach's coefficient alpha to measure reliability, and the exploratory factor analysis, which can indicate reliability when the factor loads of items measuring the construct are high, and help researchers understand the different dimensions present in the instrument they are developing (if there are more than one dimension). Instrument refining can also be performed by means of confirmatory factor analysis (Churchill's option, as he assumes that previous stages have been rigorously conducted, thus having allowed for an early formulation of the measurement instrument dimensionality).

The instrument refining stage can take researchers back to the step of generation of the set of items and alterations to the first set proposed. With a new set of items, we conduct new data collection and new instrument refining, which can be repeated until the researcher recognizes a reliable measurement that represents the construct dimensions. However, this process can be costly and involve the wasting of samples, as many data collections are not definitive. After the researcher obtains a satisfactory instrument refining, he conducts a new data collection, which is now definitive, and checks reliability again with the alpha coefficient or by dividing the instrument into two sets of items and computing the degree of association among them, or by means of test-retest reliability, which applies the instrument to the same group of respondents at two different moments and compares the results.

The definitive collection also serves for the construct validity test. To evaluate convergent validity and discriminant validity, Churchill recommends to use the Multitrait-Multimethod Matrix, which verifies association between traits (constructs) obtained with different methods; that is, with the application of measurement with different instruments, forms and/or collection moments, and even with different samples. The matrix comprised of these procedures becomes an instrument that provides comparisons between: 1) the common variation contained in a scale with several items for the same construct collected with the same method; 2) the association between the measurements of the same construct obtained with different methods; 3) the association between different constructs obtained with a common method; and 4) the association between different constructs obtained with distinct methods. The idea of making such comparisons is that, in case of a high common variation between the items for the same construct, a convergent validity is present, that is, they converge to a common measurement. This common variation should be greater than the associations of these measurements with different constructs obtained with different methods and greater than the associations between different constructs obtained with the same method.

In addition, the association between the same trait (construct) collected with different methods is expected to be greater than the association between distinct traits, either collected with the same method or not. When these conditions are fulfilled, we have evidence of discriminant validity, i.e., we actually have different measurements for distinct constructs. Pearson's correlation coefficient is commonly used to measure the proposed associations. The common variation between the construct items is usually obtained by means of factor analysis (although these measurements assume linear association, they present satisfactory results, in general).

Churchill also proposes to check the criterion validity to ensure construct validity. In a few words (we discuss this subject later), the criterion validity is checked when we have an expected and significant association between the measurement for the construct we are validating and other measurements (in general, with more consolidated operationalization) with which it may be associated in theory. If construct validity (in its several subtypes) is not achieved, Churchill proposes to restart the process from the beginning, with the construct domain specification step.

When, finally, we obtain a consistent indication of construct validity, Churchill proposes that the measurement should be presented using descriptive statistics of its distribution in the sample. The procedures proposed by Churchill and some derivations suggested in subsequent studies have been largely adopted by researchers in Marketing (for example: Netemeyer, Bearden & Sharma, 2003; Costa, 2011).

However, its rigorous application is many times unfeasible, due to the several data collections it requires, which may conflict with time and budget limitation, and the difficult of collecting data collection with distinct methods, which does not favor the use of the Multitrait-Multimethod Matrix.

3.2. An alternative to the classical model: the COARSE model

Churchill's proposal has been criticized by those who consider it too emphatic in terms of statistical adjustments for the qualitative stages of validation, and for relying almost exclusively on the concepts of coefficient alpha as the measurement of reliability and factor analysis as the measurement of validity. In addition, the procedures are for the development of multi-item scales, based on the idea that these items vary according to the latent construct variation (i.e., they a reflective relation with the construct). With such criticisms, John Rossiter developed an alternative proposal in 2002, the COARSE model, favoring the qualitative procedures in the validation of measurement instruments.

COARSE refers to the six steps the researcher should follow according to this model: Construct definition; Object classification; Attribute classification; Rater identification; Scale formation; and Enumeration. The model is presented in detail by Rossiter (2011), and we will discuss these steps below, which are a reference to improve possible limitations found in the classical Churchill's model.

The construct definition step consists in writing a definition for object, attribute and rater. The object is the focus of the measurement (for example, an advertisement). The attribute refers to what will be measured in the object (for example, the affective reactions to the advertisement); and the rater entity will evaluate the object and the attribute (for example, a sample of consumers).

The second step refers to the object classification, which involves open-ended interviews with target raters. The object can be classified as concrete singular, abstract collective or abstract formed. A concrete object is that whose meaning is known and recognized by any respondent, for example, the concept of service quality control. Abstract collective objects are heterogeneous to target raters, but they comprise a clear category to the researcher, for example, carbonated drinks (like soft drinks, flavored carbonated water or sparkling water). Abstract formed objects are those whose interpretation varies among people and have different components, for example, the concept of capitalism. If the object is classified as concrete, a single item is enough to measure it. For abstract objects, multiple items are required. In this stage, we start to write the measurement instrument items, so that they can reflect the object.

The third step is the attribute classification, also based on open-ended interviews with target raters. Attributes are classified as concrete, or formed, or eliciting. Concrete attributes are those whose interpretation is practically unanimous among raters, for example, the concept of intent to purchase. Formed attributes are abstract and characterized by the addition of a number of components that, if added following a certain combination, will form the attributes (that's why they are called formative attributes in the regular literature); one example is the concept of leadership. Eliciting attributes are also abstract, but they are internal traits of raters that can cause the responses to the measurement instrument items (which are indicators of the attribute manifestation and are called reflective indicators in the conventional literature). One example is the personal involvement of someone with a phenomenon. In the attribute classification, we continue writing the instrument items, using the single-item strategy for concrete attributes and the multi-item strategy for abstract attributes (formed and eliciting). After this step, it is possible to return to the construct definition and to include there the object and attribute components identified in the classification stages.

The fourth step refers to the identification of a rater entity, or a group of people that will judge the measurement instrument items. This step will identify the respondents in details. For this stage, experts should have evaluated the results from previous steps and participated in results improvement. In this step, we will also define whether reliability estimates are required across raters and across items of eliciting attributes.

The fifth step is the scale formation. Here, we combine the texts that contain object and attribute components to generate items. We select the scale types that will be used, with input from the open-ended interviews previously conducted with raters, and we conduct a pre-test with raters belonging to the population of interest to ensure item formulations are comprehensible. The eliciting attributes are tested for unidimensionality. Last, if the instrument has multiple items, we randomize the order of presentation, mixing the sequences of distinct attributes and objects components, so that they are not recognized by the raters, and to prevent a response pattern induced by the instrument.

The last step, enumeration, refers to: creating scale scores (aggregation strategy) based on indexes or averages; transforming them into a meaningful range for bipolar attributes, like 0-10 or -5 to +5 scores, for example,; and report the scale reliability.

Rossiter's COARSE model was well accepted for highlighting the qualitative and conceptual aspects of measurement and for expanding the range of methods beyond the factor analysis and the Cronbach's alpha, bringing the possibility to adopt single and formative indicators. However, although this proposal fueled the debate by incorporating elements not considered in the classical approach (characterized by the use of confirmatory analysis, reliability indexes and, later, structural equations modelling), the operationalization of these elements is still a challenge due to relative limitations to the methodological repertoire of researchers in Marketing, computer resources available and properties of the proposed techniques.

We understand that the power of Rossiter's argument is not exactly in the set of steps (sometimes confusing), but surely in the strong emphasis on content validity, which is evident, providing details of raters, and continued follow-up from experts for the construct of interest.

3.3. Relativizing reflectivity: the formative measurement

The current debates about measurement in our area are still about issues that appear due to the fact that the variables of interest in Marketing are usually latent and of indirect measurement. For example, fear. We know it exists, we know what it is, but we do not know how to measure someone's fear directly; we can only observe symptoms of fear in someone or ask that person to manifest it, perhaps with words or tests, how much fear he/she feels. That is, we can only observe fear indirectly by means of indicators that allow us to infer the amount of fear someone feels. In most cases, we use multiple indicators to estimate the value of a latent construct. In other cases, we believe that a single indicator is enough.

The indicators used to measure latent constructs are generally classified as reflective or formative. The reflective indicators reflect the construct intensity variation, and the formative indicators are those that, when added together, will form the construct. Some examples are provided below to better explain these two types of indicators (the first example will illustrate reflective indicators and the second example will address formative indicators).

Let's say we want to measure someone's height. We know we can measure that directly, but, for didactic purposes, let's assume we want to guess people's height with no direct measurement, only by observing people's manifestation in answers they provide to two questions. The first question may refer to the degree of difficulty the person has to reach one object on a higher shelf in a certain room. The second question may refer to the person's need to stretch or bend his/her legs while driving a car. A tall person is assumed to reach the object on a high shelf more easily than a short person, and that person probably has long legs and has to bend them while driving, whereas a short person probably has short legs and has to stretch them while driving. Thus, the answers to questions are manifestations (or symptoms) of the construct (height) and reflect the construct intensity. We also assume that, for reflecting the same construct, the answers should be correlated with one another. These characteristics turn the answers to these two questions into reflective indicators of height.

Now, let's say we want to estimate the amount of alcohol consumed by people who have just left a party, but we cannot submit these people to a blood test or use a breath analyzer. We can ask these people how many shots of whiskey or vodka or how many glasses of beer and/or other drinks they have just had. The combination of shots/glasses allows us to estimate the alcohol intake, if we know the alcohol content of individual shots/glasses. In this case, the combination of shots/glasses provides a sum that allows estimating what is not directly observable. The combined indicators will provide the alcohol rate of every person. Several independent combinations can lead to similar amounts of alcohol intake; for example, one person can drink only vodka and present the same alcohol rate as another person that drank whiskey and beer. Another person may have taken much alcohol by drinking just whiskey. Thus, the answers to different questions (amount of shots/glasses taken of every drink) should not be necessarily correlated with one another to provide the alcohol intake measurement. These characteristics turn these questions into formative indicators of alcohol intake.

Although this is a well-grounded and logic strategy, the formative measurement has had operational difficulties. Actually, although recommendations have been made for the statistical evaluation of validity and reliability (see a summary in Costa, 2011), none of them has reached the consistency of the Cronbach's alpha nor the completeness and adequacy of factor analysis. Edwards (2011) refers to this measurement strategy as fallacious and emphatically says that it should not be used.

4 PROCEDURES AND RECOMMENDATIONS

In this item, we present the main procedures and practical recommendations for the challenging task of developing and validating scales for Marketing. We particularly detail practices to access validity and reliability in multi-item reflective measurement, and such practices are the mainstream in the area. After that, we talk about procedures to evaluate validity and reliability in single-item measurement.

4.1. Validity and reliability in multi-item reflective measurement

Perhaps as a result of the broad repercussion of Churchill's article (1979), which proposed a model of reflective latent construct measurement in Marketing, and others (for example, Peter, 1981; and Gerbing & Anderson, 1988) that also dedicated attention to measurement aspects and showed failures in procedures in force in those days, we have seen, quite often over the last few decades, articles showing the use of exploratory and confirmatory factor analyses to verify the dimensional structure of variables as well as strategies to analyze convergent and discriminant validity (for example, with use of the Multitrait-Multimethod Matrix), as well as modeling by means of structural equations, among others.

However, it seems that researchers still have to analyze the question of measurement. Jarvis, Mackenzie & Podsakoff (2003), in a substantial effort to analyze the use of measurement models in Marketing, showed that the distinction between formative and reflective constructs was still confusing in scientific articles published by important journals of this area (Journal of Marketing Research, Journal of Marketing, Journal of Consumer Research and Marketing Science). From a total of 1,192 constructs used in 178 articles taken from the four journals above, 1,146 (96.1%) were modeled as reflective and 46 (3.9%) as formative. However, in the authors' opinion, 336 out of total 1,146 reflective constructs should have been modeled as formative (representing a classification error rate of 29.3%). Among the 46 constructs modeled as formative, the authors understood that 17 should have been classified as reflective (a classification error rate of 37.0%). Simulations conducted in the same study indicate the severity of such classification error, which, at most, may be the origin of errors in hypothesis testing results and, consequently, in the elaboration of final considerations of the studies.

As indicated in section 3 above, the construct nature influences the way to evaluate its reliability and validity. Considering the multi-item reflective measurement, we discussed in this section about strategies of evaluation of these aspects. Our impression when analyzing scientific articles in Marketing, especially those produced by the Brazilian academic community, is that the reports about the operational aspects of scales used to measure latent constructs prioritize reliability-related characteristics, and pay less attention to validity aspects. Perhaps this reality associates to the fact that statistical packages offer largely disseminated mathematical formulations to evaluate reliability, but less to evaluate validity. It is very important to have a clear conception that valid measurements are necessarily reliable, but that satisfactory reliability is not enough to ensure validity. We will discuss the two concepts next.

4.1.1. Reliability

According to the American Psychological Association (1985, p. 19), "reliability refers to the extent to which test scores are free from measurement errors". Pedhazur & Schmelkin (1991) classify such errors as systematic (measurement biases in the same direction in successive data collection rounds) and non-systematic (random along successive measurement rounds). For more details about error types, we recommend to read Nunnally (1978).

When addressing proprieties of estimators (in our opinion, including measurement instruments), Bussab and Morettin (2013) propose one analogy to shots from four rifles. Figure 1 illustrates the performance of each rifle.

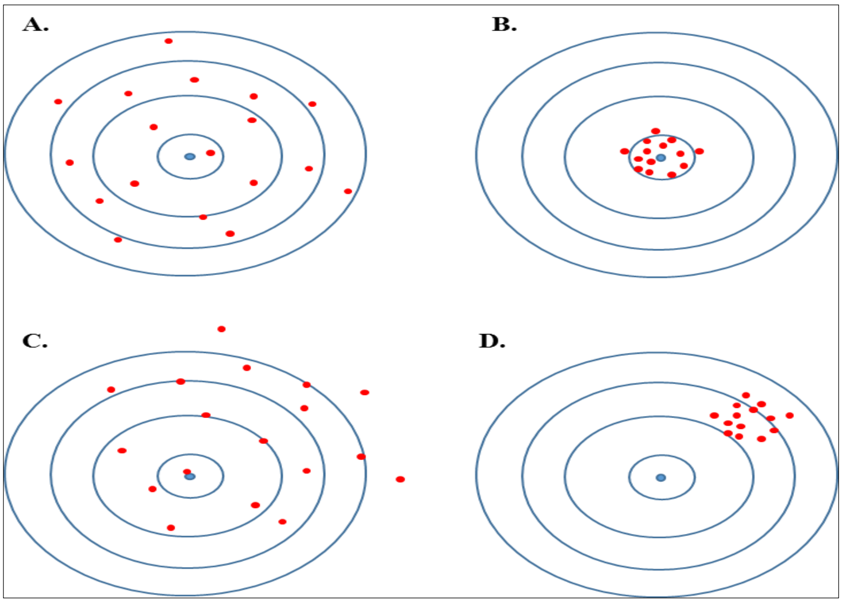

Figure 1 – Examples of estimators (bias and precision)
Source: Bussab & Morettin (2007, p. 291)

Figure 1.A shows an unbiased, but little precise, estimator (the number of shots around the target is high); Figure 1.B shows an unbiased and precise estimator (small random errors around the target); Figure 1.C shows a biased and little precise estimator; and Figure 1.D shows a precise, but biased estimator. Obviously, a desirable measurement scale should provide a score that is as close as possible to the true score, and with low variability when used repeated times (that is, 1.B).

The error type we control in reliability analysis is *precision* (consistency, scattering around a target). Let's forget for a moment the systematic component of the measurement error present in 2.C and 2.D (it will be discussed later). Using *T* for the true score (which we want to find out), *O* for observed (measured) score and *E* as the error of measurement (deviation from observed to true score), we have in reflective measurement that: $O = T + E$.

Assuming that our metrics has no systematic error, one can say that, if repeated measurement rounds are conducted, the expected value of (E) will be zero (i.e., in statistical language, $E(E)=0$). Consequently, the expected value of O will be the same as T (in statistical terms, $E(O)=T$). This is the basic idea of the **classical measurement theory**.

Then, how to access reliability in terms of *precision*? Operationally, we seek indications that the proportion of variance in a measurement attributable to the true value of a latent construct being measured is greater than the variance attributable to error components (DeVellis, 1991). Next, we discuss some examples of approaches to that.

We could think of measuring the same group of individuals twice or more times, at different moments; we would expect values obtained by the same individuals to be close, or maybe identical. Ignoring the inconvenience of contacting the same individuals at two different moments, this approach, known in the literature as the test-retest technique and usually operationalized by means of the coefficient of linear correlation between two score vectors (Pedhazur & Schmelkin, 1991), involves at least two problems: a) the carry-over effect (participating in a study may influence the answers of an individual in his/her next participation); and b) 'natural' changes in the individual's score along time (for instance, we can imagine that the level of ethnocentrism of one individual increases or decreases along his/her life). Increasing the interval between both measurements may help reduce the carry-over effect, but that may accentuate the problem of 'natural' changes, and vice versa. Evidently, these risks increase when we use multiple items to measure a construct, such as in the case of reflective measurements.

In summary, it is not an easy task to segregate reliability from temporal stability with the test-retest technique, and for this reason, we do not encourage its use in multi-item measurement (and if it is used, the interpretations should be based on the arguments presented here), although it is possible to use it in other measurement strategies, as we will discuss later.

Specifically for multi-item constructs, the specialized literature has already provided good solutions. In fact, there are efficient mathematical methods to evaluate reliability by using data from a single collection round; for instance: the Cronbach's alpha (Cronbach, 1951), the composite reliability index (Fornell & Larcker, 1981) and the exploratory factor analysis (see Aranha and Zambaldi, 2008). These methods are based on the domain sampling theory, according to which there would be several observable indicators whose variations would be caused by a common latent construct.

In order to illustrate that, let's take a reflective latent construct, intelligence, for instance. Assuming that we will reach a consensus about its concept definition, we can imagine characteristics of individuals based on which we could define it. One example would be the time required to solve problems. We will state that more intelligent individuals solve problems more quickly. If we elaborate a measurement instrument with ten types of problems and these problems are solved by, let's say, 300 individuals, we will expect the times to solve each type of problem are positively correlated (the more dependent on the intelligence variation is the variation of these times, the better the measurement).

Despite the Cronbach's alpha limitations (for example, if the other aspects are unaltered, the greater the number of similar items and the number of items in a scale, the greater the value, and, especially, the fact that a high value for the measurement does not *ensure* construct unidimensionality), its use is justifiable in the reliability evaluation of a scale, in particular in an early state of item refining. We interpret low alpha values (there is no consensus on a minimum acceptable value; but we recommend at least 0.60) as indicators of low internal consistency and, consequently, the items are useless, requiring the elaboration of new ones or adaptation of existing indicators (it should be noted that, when our construct is formative, there is no sense in expecting a high Cronbach's alpha, as the correlation among the items is not a requirement).

As an alternative to the coefficient alpha to measure reliability, we can use the composite reliability index proposed by Fornell & Larcker (1981). Composite reliability can be obtained with a factor analysis and it indicates the proportion of variance of the true score of a construct in relation to the total variance of the calculated score. For not involving the inconvenience of inflating it with the inclusion of items that are similar to the others in the scale, its use has become popular and we recommend the composite reliability index, rather than the Cronbach's alpha. However, the composite reliability index cannot ensure construct unidimensionality either. Just as for the coefficient alpha, we also consider as reliable, constructs presenting composite reliability over 0.60.

Regarding the exploratory factor analysis, we should expect high factor loads (at least 0.40 or 0.50; we point out that there is no minimum value established by consensus) between the indicators and the factor that represents the dimension to which they would belong to[5].

It is possible, for instance, when including several items with similar text in a scale, to inflate their internal consistency indexes. However, it does not make the measurement instrument more effective, it takes up space in questionnaires and makes them unnecessarily longer. In this sense, precautions should be taken in the item generation stage to capture complementary aspects of the same construct. We recommend the article by Lee & Hooley (2005) about theoretical fundamentals, applications and limitations of coefficient alpha and factor analysis, and the article by Costa (2011) about the scale item development stages.

[5] In our perception, in general, when an exploratory factor analysis is reported in articles in Marketing elaborated by the Brazilian academic community, procedures of orthogonal rotation are used (which involve null linear correlation between the factors extracted). However, it seems reasonable to assume that it is common that dimensions of the same reflective construct are correlated (when dealing with multidimensional constructs). For this reason, we understand that the proper rotation procedure would be oblique (for a more specific coverage of the subject, we recommend the study conducted by Stewart (1981)). It should be noted that, in formative indicators, we should not necessarily expect the high factor loads mentioned above.

4.1.2 Validity

We understand the validity of a construct measurement is the extent to which the proposed measurement actually evaluates what it purports to measure. The possible presence of systematic errors (see previous item) should be captured when employing effective validation procedures. It should be noted that, initially, we have to accumulate 'evidences' that our measurement instrument is valid; it is not possible to *be absolutely sure* that validity will be reached, as it would require the latent construct being measured to be observable.

Our objective, when seeking evidences of validity for a scale, is to provide reasonable conditions of construct measurement, so we can test assumptions involving the construct. Unlike the methods to evaluate reliability, the methods available to evaluate validity depend on the researcher's skills to develop more or less efficient strategies. These strategies can evaluate three types of validity[6]:

a) **translation** – it is a non-statistical and qualitative type of validation that involves the systematic analysis of measurement instrument content to evaluate whether their components represent the construct facts properly (and in this case, we say that content validity is present) and whether the text and form are adequate to be applied to the target population (and in this case, we say face validity is present). In general, this type of validation is conducted by experts (researchers or participants); and it is possible to use potential raters as judges too;

b) **criterion** – it involves the analysis of the foreseen association between our measurement and a variable taken as a criterion, representative of the construct. For example, the measurements of the bias in a scale of donating behavior can be compared to the donating behavior, evaluated in the following year. The criterion validity, in this case, is qualified as *predictive*. It is possible to employ a criterion validity, for instance, when measuring materialism among religious people and among business students, just as conducted by Belk (1985);

c) **construct** – it refers to the degree to which the operationalization of a construct shows to be adherent to the theory, regarding its definition and properties. Its dimension structure and its relationships with the other constructs are also evaluated. Validity subtypes are: convergent, discriminant, nomological, and known-group validity. Here, the associations found between the construct and others are confronted with the theoretical expectations, and techniques such as the Multitrait-Multimethod Matrix, the Confirmatory Factor Analysis (CFA) and the Structural Equation Modeling (SEM) are useful for such a verification.

We should see the types of validity analysis strategies as complementary. We rarely find, in articles related to Marketing, the simultaneous use of all types.

To illustrate how these strategies are applied, let's take the example of Richins & Dawson (1992). These authors, when developing and proposing a widely used scale to measure materialism, used strategies of criterion validity. In questionnaires sent to raters, besides including the indicators of the scale of material values, they also presented questions like: what is the income level required to fulfill your needs?; what is the relative importance of values such as financial security, pleasant relationship with others and self-actualization?; what would the rater do if he/she unexpectedly won a certain amount of money (egoist or altruist use)?; among others. Then, they used a solid theoretical base to justify behaviors expected from groups of more materialistic and less materialistic individuals, and they analyzed if the score of material values indicated by the proposed measurement instrument could predict the behavior in the rating questions presented. It should be noted, in this example, the effort of considering about the characteristics expected for groups of more and less materialistic, and of creating protocols to seek validation.

There are many techniques of validity analysis in each strategy, and detailing them is not within the scope of this article. We can attest that the classical methods of validity evaluation using these strategies are well documented (cf. DeVellis, 1991; Netemeyer, Bearden, & Sharma, 2003; Costa 2011). However, the use of more sophisticated statistical techniques for validity analysis has been intensified lately. For example, Gonçalves (2013) uses a model of third-order confirmatory factor analysis to evaluate reliability and convergent validity for a satisfaction scale. This construct was defined as having three primary dimensions – service core, peripheral aspects of service quality and value. The dimension of peripheral aspects of service quality has, in turn, three subdimensions and the value dimension has two subdimensions.

Yi & Gong (2013) proposed to measure the behavior of consumer value co-creation with a hierarchical and multidimensional approach. The strategies of (convergent, discriminant and nomological) validity used by the authors involved first and third-order confirmatory factor analysis

[6] Although the focus of this item (4.1) is on multi-item reflective measurement, these strategies of validity analysis are applicable to other measurement alternatives, as discussed later. The variation of applications is in the techniques employed.

models and the PLS (Partial Least Squares) model.

4.2. Validity and reliability in single-item measurement

The strategy of multiple item measurement, which involves the application of relevant techniques (such as structural equation modeling), assumes that a well-delineated construct is measured from the analysis of scores for two or more items. In this perspective, and as indicated above, every item measures a construct facet, which, according to the domain sampling theory, is directly associated with the construct for having some of its variation derived from the latent factor variation (the remaining portion of its variation is explained by a random error). However, there is an alternative measurement frequently used in research in Marketing, which is the measurement of constructs with a single item, instead of multiple items.

The central idea of the domain sampling theory makes the statistical validation of construct or dimension measurements very easy. In fact, if we consider that the content and face validity of a set of items is good (this state is more qualitative), the statistical validity is easily analyzed by using the analysis of factor adequacy and internal consistency. On the other hand, in a single-item measurement, there is no sense in submitting it to a factor analysis or internal consistency analysis, using, for instance, the Cronbach's alpha or the composite reliability index. That forces the use of distinct techniques of validity analysis. We present below the main procedures for validity analysis, considering first the qualitative evaluation, and then the alternatives for statistical evaluation[7].

4.2.1 Qualitative stage of validation

In qualitative evaluation, for this type of scale, the precautions are the same as for multi-item scales, and the goal is simple: make the item presentation reflect the construct content expressed in the construct definition. Besides the clear association with the definition, that is, the content validity, and to ensure a good face validity, the item presentation has to be concise and understandable, even if the scale is smaller (compared to the multi-item measurement). In other words, the fact that the measurement is based on a single-item does not imply the use of a very extensive item or a presentation that is not suitable to the rater's understanding, even if the measured construct is abstract. That brings an even greater challenge to the researcher, considering the need to consolidate in a single item presentation all the meaning of a construct, besides the requirement of a presentation for that is coherent with the rating scale to be used.

As a method for this challenge, two procedures have to be carefully used: first, the item should be elaborated and submitted to the evaluation of experts on the theme and/or experienced researchers; second, the item should be exposed to future potential researchers, to evaluate their understanding of the association between the concept and the item. These procedures help ensure content validity (association between the item and the definition) and face validity (item presentation and understandability).

John Rossiter (2011), in his COARSE model, emphatically states that the qualitative stage of a single-item measurement is the main, or perhaps the only, way to ensure the validity of a scale. Yet, we understand that the reiterated indication of content and face validity by experts or potential raters of the scale is not enough, or at least that there is no loss when confronting it to results of a concrete application of the scale for the construct it purports to measure.

4.2.2. Quantitative stage of validation and reliability

The consistency analysis of a single-item scale is confirmed with data originated from the evaluation of the **adherence of sample results to the expected behavior** of the variable that gave origin to the sampling, of the **criterion validity**, of the **known-group validity** and of the **test-retest** procedure. Let's see some details and recommendations.

Regarding the **scale adherence to the expected behavior**, let's assume that the metrics should measure a construct whose measurement follows a certain probability distribution in a population. For example, it is possible to say that the 'level of satisfaction of the population with the government' presents a symmetrical distribution like a nearly normal behavior, or that the 'level of propensity of young people to civic participation' is asymmetrical to the right, with greater concentration in lower scores of a scale. In these terms, if we apply a scale to measure these constructs, the behavior of sample scores should reflect approximately the expected distribution model.

In an operational perspective, this analysis can be conducted in an exploratory manner or with tests, but we recommend a well-grounded exploratory evaluation. For instance, an evaluation of a histogram or a stem-and-leaf plot of sample values may be enough to indicate if the sample format matches the expected distribution. Naturally, it is not always possible to assume a distribution for the reference variable, which makes this type of analysis more difficult.

[7] Considering the purpose of this article, which is to be used as a reference to researchers, and the less traditional use and development in Marketing literature, we decided to detail these procedures and provide more recommendations, unlike item 4.1, for which the theoretical and application development is much broader.

We also recommend the evaluation of the variable behavior in relation to some statistical measurements. For example, the scale is almost always expected to capture the actual variation of construct intensity that exists in the universe of interest. Thus, if in a population with moderate dispersion in the construct intensity, a scale generates a very low and a very high standard deviation, it may indicate adequacy problems of the metrics that do not allow it to capture the expected data behavior.

Regarding validity analysis, unlike the multi-item measurement, we recommend only two procedures for single-item measurement: the criterion validity and the known-group validity. In the **criterion validity**, the procedure will analyze the scale behavior in the prediction or association of the construct being measured in relation to another construct with a previously validated scale (when we expect such prediction or association). For instance, let's say we are analyzing a single-item scale to measure the 'declared level of environmental awareness', stated as follows: 'I'm aware of environmental issues' (to be evaluated in an scale of agreement); if we know the environmental awareness is a predictor of the 'tendency to buy products with a sustainability certificate', and if we already have a previously validated scale for this construct, then we can easily see if our current scale is valid or not, by applying the two metrics simultaneously, and checking for the expected association, that is, checking if we observe a significant correlation between the measurements of the two constructs, or if a regression analysis can reach proper levels of adequacy (according to what we expect in terms of prediction intensity and direction).

Almost like the criterion validity, it is also possible to analyze the expected behavior of a measurement in relation to groups or specific categorical variables, in the **known-group validity** (this strategy is not frequently used in multi-item measurement). For example, in a single-item scale to measure 'trust in city administrators', stated as 'in general, I trust the administrators of my city' (with evaluation in an agreement scale), and if we know that people with a link with the leaders have a more positive evaluation than people without such a link, then the scale will be valid if it properly reflects this difference. That can be evaluated, for instance, using statistical techniques like variance analysis or Student's t-test, or corresponding non-parametric techniques (Kruskal-Wallis test or Wilcoxon-Mann-Whiteney test). Thus, if data behavior is as expected with the indication of these tests, it is possible to ensure, or not, the known-group validity.

Last, as one way to evaluate reliability, we can evaluate single-item scales by means of its behavior at different moments of application in time, by using the **test-retest** procedure[8]. We apply the scale to a group of raters at a certain moment in time, and later, we conduct the second application to the same group, after a short time, enough to ensure the construct intensity will not vary too much, and considering enough time distance that will not allow raters to remember their previous answers. Reliability is ensured if data correlation at both applications is sufficiently high to reflect the expected behavior of behavior convergence (we recommend at least 0.8).

Table 1 summarizes the procedures we recommend.

Table 1 – Validation procedures for single-item scales

EVALUATION	RECOMMENDATION
Content and face validity	Exposure of scale to experts and potential raters and qualitative evaluation of results
Performance adequacy	Analysis of measurement and behavior (distribution) of sample data in comparison to the behavior we expect
Criterion validity	Analysis of the scale association or prediction in relation to other constructs with scales previously validated and comparison to the results we expect
Known-group validity	Analysis of scale measurements in relation to groups of individuals and comparison to the results we expect
Test-retest reliability	Evaluation of the association between the measurements generated by the scale at two distinct moments in time and comparison to the association we expect

[8] As indicated above, we do not recommend this procedure for multi-item measurement, as we have consistent methods for just one evaluation. This is not the case of a single-item evaluation, and then, that's why this procedure is useful here.

4.3 Relevant Complements: dimensionality, instrument organization and common-method variance

A relevant question in current debates about measurement in Marketing involves the dimensionality of a construct. A construct does not have to be necessarily unidimensional, since it may have several dimensions (subconstructs) or attributes (according to Rossiter). For instance, let's take 'trust', a construct that can have, according to the literature, multiple dimensions, such as perception of honesty, benevolence and competence. In this case, we understand that, to measure trust, it would be necessary to measure these three dimensions, that is, if raters consider the object under analysis as honest, benevolent and competent. The three dimensions, or attributes, can also be abstract and, in this case, require multiple items to be measured. Confirmatory factor analysis is a useful technique for the analysis of instrument dimensionality (see Aranha & Zambaldi, 2008), but it is limited to adjustments of reflective models. It should be noted that the dimensionality test of a scale should not be based on the alpha coefficient, the composite reliability index, nor exploratory factor analysis, but on more robust procedures instead.

Besides the concerns about qualitative and quantitative procedures for the construction and validation of measurement instruments, we have other concerns related to their application. In this domain, we include the collection form (for example, interviews or self-completion), the application moments and the distinct samples to which the instruments can be applied. Each variation in application is subject to bias and, when such bias has great influence on data, we have an undesirable phenomenon, known as common-method variance, which is a common pattern to all (or most) answers from raters, either due to their socially desirable behavior, or to their attempt to guess what should be measured and to influence it, to their struggling to seem coherent, or to biases inherent to the data collection process (such as poor understanding of an item or any type of induction from the interviewer).

The use of multiple methods to collect data of a construct to mitigate the common-method variance takes a long time and requires many other resources and, for this reason, the researchers, facing difficulties to use tools like the Multitrait-Multimethod Matrix, employ techniques to minimize the potential bias resulting from the use of a single method. One way to make it difficult for raters to recognize what is intended to measure may be mixing the presentation order of items from the dimensions present in the instrument, as we already mentioned when presenting the proposal of Rossiter (2002, 2011). Another way would be to use inverse items (those with negative conceptual relation with the construct) among items with positive relation with the construct (see Wong, Rindfleisch & Burroughs, 2003; and Aranha & Zambaldi, 2008). For instance, to measure competence, we could include statements that refer to such attribute, along with an item that refers of incompetence. The presence of inverted items tends to force raters to focus on their answers, as they cannot adopt an automatic pattern of answering (such as a strong agreement with all items). Evidently, inverse items should have inverted values for score computation and analysis. In addition, it is difficult to elaborate such inverse items, as they usually have negative sentences, which may confuse raters.

5 CONSIDERATIONS ABOUT RATING SCALES

One important aspect of construct measurement in Marketing is the "rating scale", which is associated with the reference that raters have when they select the number that will indicate the construct measurement. In fact, when raters select a level of intensity, they usually do it by indicating a number chosen among a set of options (for example, 5 scores numbered 1 to 5 in an agreement scale). It is always a good challenge for researchers to define proper numeration alternatives to their different research purposes.

Rossiter (2011) says that we obtain the validity of a scale by adding item content validity (a statement to capture agreement, for example) to rating scale validity (or the number of scores and the meaning they have for raters). It is easy to agree with Rossiter, what requires special attention about the measurement decisions.

We present here the main decisions to be made and the most adequate alternatives to each context. In general, these decisions are related to **rating scale coherence** with the item presentation; **number of scale scores**; the **aggregation strategy**; and the alternatives of **statistical operationalization**.

5.1 Coherence between scale and content presentation

Regarding the coherence between scale and content, the idea is to ensure the rating scale coherence with the item presentation. For example, if a researcher decides to present an item as a statement, the rating scale makes sense if it is an agreement scale with the statement, with different levels. That is the most frequent type in measurement in Marketing, the 'Likert scale', proposed by Rensis Likert (1932). The frequent problem observed here is when the rating scale is in the form of an agreement scale, but without the item presented as a statement (to which raters should indicate if they agree or not, and at what level).

There is no sense, for instance, in asking a user to evaluate a service (for example, “evaluate the public transportation service quality”), and right after that include a 5-score agreement scale (for example, 1 for totally disagree to 5 for totally agree). Evidently, to avoid this type of problem, the coherence between the answer alternatives and the way the item is presented should be carefully analyzed by the researcher, who should also obtain directions from experts and potential respondents.

Also regarding the coherence between the rating scale and item presentation, one aspect that has not been highlighted in scale development refers to item valence. This problem appears especially when the measurement analyzed involves attitudes. According to their nature, attitudes are associated with general evaluations, which, in most cases, vary from a negative to a positive meaning. That is, the indication of a measurement related to attitude may bring two findings at the same time: first, if that is a positive or negative evaluation; second, the magnitude in either option (that is, if negative, how negative, and if positive, how positive).

Rossiter (2011) suggests that, for an evaluation construct, or, in general, a ‘bipolar’ construct, the most coherent rating scale is that with alternatives of negative, null and positive values. Using the example above, a measurement to evaluate the transportation service quality, with an item presentation as follows: “evaluate the public transportation service quality”, the most coherent answer alternatives would be (using a 5-score scale), -2, -1, 0, +1, +2. Another option could be provided, for raters to select one score in a 11-point scale, from 0 to 10 (0, 1, 2, ..., 10) or from 1 to 100 (leaving space for raters to indicate a number between 0 and 100). We understand that this decision does not define something right or wrong, but something ‘more adequate’ to every context.

5.2 Number of point in a scale

As it is a simple practice, most metrics in Marketing use measurement scales with intervals (with a minimum and a maximum point) (see Stevens, 1946) and with a limited number of alternatives (for example, a 7-point scale, with 1 indicating the minimum magnitude and 7 the maximum magnitude). The great advantage of this decision is related to the safe generation of answers and easiness to raters. The great disadvantage is that some statistical techniques cannot be used or adapted.

Regarding this aspect, we know the main statistical techniques used in studies in Marketing assume that some distributions are continuous. This is the case, for instance, of the classical technique of normal linear regression, which, for assuming model error, requires the variable answer to be continuous. The way we usually conduct analyses makes it difficult to accept continuous answers, as we use discrete measurement, limited to a certain number of points.

In fact, there is no rule for the definition of the number of points, but we can say that the scale should have as many points as possible. Actually, if it is possible to ensure raters the possibility to indicate a number, the researcher would only have to define scale limits, what would ensure a sense of continuity in the measure; this would allow us to use many statistical techniques without big obstacles. However, this alternative involves operational restrictions: as most researches use questionnaires, the indication of a number by the respondent would make it difficult to collect answers, e possibly generate a large number of missing values.

On the other hand, there are boundaries to be considered. We understand the size of questionnaires is the main aspect that defines the number of points, requiring us to consider that many points tend to occupy more space and that may produce very long questionnaires and affect answers. In addition, the capability of raters to provide a reliable answer with a certain number of points should be taken into account. This aspect is especially relevant for cases in which raters need denominations about the points, that is, the intensity indication for every scale point. For instance, in a 5-point scale, it is easy to denominate points as: 1 – I fully disagree; 2 – I partially disagree; 3 – I agree/disagree moderately; 4 – I partially agree; 5 – I fully agree. On the other hand, in a 11-point scale (from 1 to 11), it is very complicated to provide intensity indication to every score[9].

Our recommendation is the following: if the space is enough, we should use as many points as possible, preventing, however, specific denominations to every score. An interesting strategy seems to be to use 10- or 11-point scales (from 1 to 10, or from 1 to 11, or -5 to +5), denominating only the extremes and with an indication of meaning in the intermediate region (see Hodge & Gillespie, 2007). Applications with this type of scale have been considered consistent, and, to an extent, facilitate answers, as in the Brazilian culture we commonly use 0 or 1 to 10 points (see Barboza *et al*. 2013).

[9] The denomination of points is more complicated in cases of odd numbers of scores, because there is a tendency to associate the central point with the condition of an indifferent or neutral score, which effectively has no meaning, as the indifferent or neutral score provides no answer in the scale (for example, a neutral person in the agreement with a certain statement actually does not assign any score in a scale that measures exactly the level of agreement).

The option for using as many points as possible is, however, controversial, and it depends on the rater's capability to understand how the scale works. According to our field experience, especially with raters with low levels of education, reducing the number of options can be useful, as it simplifies the answer indication. We can use warming-up items, for example: 'It's cold today.' or 'I like soccer', to check for the understanding of how to manifest agreement with the items to be read. That is possible when the application is made by an interviewer, in person.

5.3 Aggregation strategy

The problem of aggregation appears when we use a multi-item scale to measure a certain construct or dimension. The demand comes from the need to access the total construct value (sometimes, this measurement is not necessary, for instance, in studies that test models through structural equation modeling). For cases requiring aggregation, we highlight here three options for constructs with reflective measurement and one option for the other cases.

If we have a set of items that reflectively measure one construct, and if such set of items is adequate in terms of factor structure and internal consistency, the first and the most common recommendation is to use the aggregation strategies from factor analysis, which is included in most computer applications. That is, in factor extraction, we rely on the software to generate a total factor measurement. The problem of this strategy is that, in currently implemented routines, the variable generated is standardized in such way that its average is 0 and its variance is 1, which normally differs from the measurements of item origin scales (between 1 and 5 or 1 and 7, among others).

For this reason, if the psychometric structure is adequate, it is possible to keep the measurement aggregated in the same scale of variables by extracting simple arithmetic mean values of every rater in the set of items (that is, extracting the scores mean values of every rater) (Bagozzi & Edwards, 1998) or, by extracting a weighted mean of scores by rater, using the factor loadings of the respective items as the weighting criteria. This second strategy has the advantage of keeping the scale aggregated within the limits of the original scales, providing greater weight to items with better correlation with the latent construct (reminding that the factor loading is, under some premises, a measurement of correlation between the variable and the latent factor).

If a construct is measured with a multi-item scale, but without a reflective relation, the best aggregation strategy is the extraction of a weighted measurement by rater. Here, we have the need to justify the weighting factors; otherwise, any aggregation is risky. The aggregation with simple arithmetic means of scores by rater is possible in case of total absence of a weighing reference, but the measurement analyses have to consider possible problems resulting from this procedure.

5.4. Statistical operationalization

We present here brief considerations about the statistical operationalization of data from commonly used scales. This subject is controversial, depending on the researcher and his/her level of theoretical requirements. For this reason, we will limit our considerations to some evaluations and recommendations of practices, which are subject to contestation.

As discussed above, several techniques involve continuity of variables in their application, such as some conventional linear models. For this reason, if we are operationalizing data from scales that use a certain number of scores, such data will hardly have a similar behavior to that of a continuous variable. That does not allow us to use, for example, the techniques of multiple regression from the normal linear model (and even some of the techniques of general linear model, like quantile regression and others) when the variable answer is measured using a Likert scale, for instance[10].

We understand that the most coherent solution for this situation is to increase the number of techniques, using as many as possible, and to analyze convergences, similarities, analogies and discrepancies, thus allowing to build a complete reference about the reality under study by using all options available (Haig, 2005). That procedure uses complementary techniques to commonly used methods (which, in our opinion, can be applied as well, provided that they are properly considered in the results evaluation). That is, we understand that it is possible to use conventional techniques, as well as parametric methods of prediction/association for discrete data (present in general linear models and categorized data analysis models, for instance; see Faraway, 2006, Sheather, 2009), and non-parametric or semi-parametric techniques (see Kloke; McKean, 2013; Hao, Naiman, 2007).

Table 2 summarizes the procedures we indicate in this section.

[10] This statement is controversial, as it often mixes variable continuity with scale continuity, which effectively are distinct concepts. Either way, it is not difficult to find applications of models that use continuity with variables measured using discrete scales (and likewise we find several applications of parametric techniques without total security regarding the distribution concepts involved).

Table 2 - Procedures for rating scales

EVALUATION	RECOMMENDATION
Scale-content coherence	Analyze carefully the association between item content and numerical alternatives offered to raters and indicate proper numbers for item meaning.
Number of points	Use as many as possible, considering the questionnaire space and easiness to raters.
Aggregation strategy	For multiple items, if the measurement is reflective, confirm the psychometric consistency and aggregate by means of factor analysis or of scores by rater, either through simple arithmetic mean or weighted mean by factor loadings.
Statistical technique	Conduct complementary analyses involving classical techniques and other parametric techniques, as well as non-parametric and semi-parametric methods.

6 CONSIDERATIONS ABOUT MEASUREMENT IN CROSS-CULTURAL RESEARCH

Cross-cultural studies have become usual in social sciences and are conducted to test the generalization of theories or to provide a ‘natural’ experimental treatment to study the influence of culture on behavior. Alternatively, we may think that a study on a single culture can lead to a partial view of reality or to a (wrong) generalization of results from one culture as if they were universal (Steenkamp, 2005).

In Psychology, it is common to see efforts to access universal dimensions of personality, such as values, beliefs and emotions; however, cultural systems may shape these individual characteristics differently. According to Church (2010), the existence of universal dimensions of individual differences that can be accessed regardless of the context – and in similar manners across cultures – has been questioned.

In particular, in the Brazilian academic community in Marketing, versions of scales (with varied degrees of adaptation) developed in other countries are frequently used. Besides the attention to the aspects of measurement reliability and validity when applying measurements to a different context from the one it was developed for, and in particular when the researcher intends to make cross-cultural comparisons, other types of interference should be observed. Van de Vijver & Leung (1997) sort these interferences into three groups: construct bias, method bias and item bias.

Construct bias occurs when the definitions of a construct have partial overlapping of cultures. In these cases, we say that there is no conceptual equivalence. Church (2010) gives an example, the achievement motivation concept, which can be more socially oriented – emphasizing goals of social or family groups – in collectivist cultures in the comparison to the Western conception, which emphasizes individual efforts to achieve personal goals. Milfont & Fischer (2010) presented a literature revision of measurement equivalence and one step-by-step model of rating by using a confirmatory factor analysis.

Method bias can assume three forms (Church, 2010): (i) sample bias; (ii) instrument bias; and (iii) administration bias. One example of sample bias can occur in an investigation of individuals belonging to a certain socioeconomic group. What would be the equivalence criterion among Brazilian, North American and Japanese raters, for example? The use of a socioeconomic classification criterion that matches the Brazilian reality, as proposed by Kamakura & Mazzon (2013) is (probably) not directly applied to the reality of these other two countries. Would income or buying power be superior approaches to identify equivalence? Similar criticisms can be made to studies that seek to measure the poverty of nations (would there be a universal criterion of poverty or is that a concept that should consider regional specificities?).

Instrument bias refers to the difference in the interpretation of data collection instrument by raters, for instance, resulting from question writing. Wong, Rindfleisch & Burroughs (2003) indicated problems with the administration of items directly written among raters from East Asia; they claim that, due to a greater tendency to agree with sentences made by third parties, items written as questions could better capture values. Reardon & Miller (2012) suggest that benefits can be obtained with the use of metaphors in scales, when comparing it to the use of more traditional forms, such as Likert and semantic differential. Administration bias refers to the difficulty found in the communication between researchers and raters.

Item bias occurs when individuals with the same intensity of a characteristic, but belonging to different cultural groups, display different probabilities of answering items in an expected direction. Regarding linguistic equivalence, the procedure of reverse translation is probably more used in Brazilian studies, but there are others available. For instance, it is possible to manage an instrument in two languages for bilingual people and compare the correlation between the answers.

With increasing globalization of science and societies, cross-cultural studies will probably be more and more important, as well as the need to successfully address unresolved issues of measurement in theses contexts. According to Church (2010), valid measurements between cultures will require continuous developments of researchers in statistical methods to determine measurement equivalence. For example, hierarchical linear models and their ability to simultaneously test assumptions at both individual and cultural levels will probably be more and more important.

7 MEASUREMENT TRENDS IN MARKETING

This section shows some topics of measurement trends in Marketing. Our selection was based on an evaluation of recent literature related to studies and measurement in Marketing. The reference themes were: the Item Response Theory (IRT), Bayesian estimators and Partial Least Squares (PLS) modeling.

7.1 Item Response Theory

According to Church (2010), the item response theory (IRT) has been used to measure a number of latent constructs, such as intelligence, personality traits, individualism and collectivism. It has been used for more than 60 years, more frequently in the fields of Education and Psychology (Samartini, 2006), with applications also in Brazilian studies in Marketing (see Lucian, 2012).

Although the IRT represents a set of models with varied specificities, most of them have two parameters in common. The first refers to the extent the item (question) is close to the trait to be measured; and the second refers to the extent the trait is present in the rater (a third parameter associated with randomness could be considered, depending on the study). Researchers in Education have developed several studies using the IRT, by using the item parameter to measure the difficulty of questions in an evaluation and the rater parameter to measure the skill (or knowledge) of students. This technique has become popular as a way to standardize results from students that are submitted to different evaluations, thus allowing their performances to be compared.

The field of Psychology, traditionally involved in the measurement of latent traits, also presents a large collection of IRT applications when seeking to quantify the adherence of items of an instrument to the construct to be measured and the presence of the construct in raters. In the fields of Business Administration, such as Marketing, IRT applications are still less common, with prevalence of classical approaches such as factor analysis (FA) and structural equation modeling (SEM).

However, there is a tendency of increasing the use of IRT in Business Administration and, particularly, in Marketing, encouraged by some of its properties that allow more information and more stable results than in classical approaches. One of the advantages of using the IRT is that, when a measurement of the extent of the trait presence in the item is obtained, only a few questions are enough to identify the item intensity in the rater. That is possible because IRT models provide a distribution of probabilities of the possible answers for each question in relation to the level of the trait presence in the rater.

The IRT includes a number of different models, which can collect binary or scalar data (Scherbaum, Finlinson, Barden & Tamanini, 2006). Developments in terms of tools and applications have been greater for binary data and, for this reason, we believe the use of IRT in Marketing (a field that tends to use multi-item scales), although increasing, is still incipient and tends to remain this way in the medium term.

The IRT models can be sorted as cumulative and unfolding models (Samartini, 2006; Scherbaum *et al*., 2006). Cumulative models assume that possible answers to an item imply order and that any progress of such order increases the extent of trait presence. The agreement scales, in this context, would indicate that the more a rater agrees with a statement (not an inverse statement), the more it will have from the trait. Unfolding models, in turn, do not assume trait accumulation in the order of possible answers to an item. Let’s take, for instance, the following statement: 'smoking should only be allowed in open areas'. Someone that absolutely agrees with smoking in any area would fully disagree with this statement, just as someone that absolutely disagrees with smoking in any area would. People who are not at the extremes of opinion about the permission to smoke would choose their answers at intermediate levels of agreement. As unfolding models do not assume trait accumulation according to an order of answers to items, they bring a distribution of probabilities to each possible answer in relation to the presence of trait in every rater.

The comparison of results from the IRT and the confirmatory factor analysis, obtained either by simulations or empiric studies has demonstrated greater adequacy of IRT (see Salzberger & Koller, 2013; and Buchbinder, Goldszmidt & Parente, 2012) in measurement validation. Apparently, measurements validated by IRT are more stable in distinct contexts, while measurements validated by AFC require more adaptations in when the contexts vary (we understand distinct contexts as variations in data collection forms

– for instance, personal interviews or via telephone, and self-completion of questionnaires –, collection moments, and samples that represent distinct populations (Meade & Lautenschlager, 2004)).

The theoretical properties of IRT can explain these differences. One of them is that the characteristics of items and individuals obtained from the answers provided are independent of one another. In other words, it is possible to determine parameters of items (question difficulty or trait presence) based on different sets of raters representing distinct populations (Salzberger & Koller, 2013; Scherbaum *et al.*, 2006).

The classical models (factor analysis and internal consistency analysis) are based on the correction of construct scores and estimate parameters to items. For not considering that the parameters of items are separated from the parameters of raters, the results are limited to sample characteristics and, consequently, to sample representativeness. This is one of the probable reasons why adaptations are required to scales already validated by a classical approach in a culture when cross-cultural studies are conducted, as the scores are limited to the original sample and the parameters of items are dependent on it, which does not occur in the IRT, at least in theory. This bias in the classical approach is one disadvantage in terms of result stability. Another advantage of the IRT is that the standard error of items varies at all trait levels, that is, it is possible to determine the latent trait for each of its level (Scherbaum *et al.*, 2006).

However, unlike classical approaches, the IRT models do not measure reliability of a complete instrument of measurement when using multi-item scales, as reliability in the IRT is evaluated by item (Scherbaum *et al.,* 2006) and do not offer general measurements, such as the composite reliability index, and this can be considered a disadvantage.

Another disadvantage of the IRT is that it requires larger samples than in classical approaches (Church, 2010; Scherbaum et al., 2006). In addition, the use of IRT is complex for users not familiar with advanced statistical methods and, due to lack of computer resources in terms of software with user-friendly interfaces of IRT, we believe that its use is and will be inhibited in areas other than Education and Psychology, where more prominent developments have been conducted.

The IRT models also present two concepts and at least one of them can be considered a disadvantage in relation to classical approaches, which is the trait unidimensionality. The IRT models usually assume that one instrument measures a single trait, although multidimensional models of IRT are available, but they are very complex and presume difficult implementation (Buchbinder *et al.,* 2012; McDonald, 2010). The models of classical approach are more easily adjusted to multiple traits in instrument validation.

The other concept of the IRT models is local (or conditional) independency, which means that the answers provided to one item exclusively depend on the latent trait and do not affect nor are affected by the answers to other items. This concept can explain the fact that researchers who prefer the IRT claim that the parameters of items do not depend on the sample and then because of that the estimates are stable. However, this argument is questionable, as it refers to a concept that is not always observable.

The comparison between the properties of IRT and classical approaches allows us to imagine situations for which the selection of one approach or another is more or less adequate. We could recommend the classical approach to situations with presume constant standard error at all levels of a trait in one item. However, such situations are not very plausible, which favors the selection of an IRT model. IRT is also the best choice when there are no representative samples of the population for which the measurement will be developed available, due to its independency between the parameters of items and parameters of raters. The same property, for ensuring greater stability of item parameters regardless of the context, also places the IRT as the first option to create new measurements or refine existing measurements.

When the purpose is to evaluate the general reliability of an instrument, we could recommend a classical approach, while IRT would be more adequate to obtain reliability at the different levels of the trait for each item. Another criterion to be considered may be parsimony, also questionable. Whereas unidimensionality in IRT assumes a more parsimonious model, on the other hand, it restricts multidimensional models that can make better sense in some theoretical formulations. In addition, IRT models involve complex application and require more computer and technical resources than classical models, particularly when we adjust multidimensional models, which makes IRT use less parsimonious.

One good application of the IRT is the case of extreme response style, or situations in which answers are at the extremes of questions (of Jong, Steenkamp, Fox & Baumgartner, 2008). As decomposition models present distinct probabilities to each answer in relation to the trait value of every individual without presuming trait accumulation according to the order of possible answers, it is possible to better discriminate the trait in raters that select full agreement or full disagreement with a certain item than in cumulative models. Besides this attribute, there is a possibility to have variation in the standard error of the trait at every level, thus allowing for different degrees of precision for individuals at extremes or at intermediate levels of the trait. The capability to properly handle the extreme response style is also a benefit of the IRT in the treatment of common-method variance – the rater’s tendency to have a unidirectional position (either very favorable or very unfavorable to the trait along his/her answers) across the measurement instrument. When allowing variation in the estimate of item parameters for different trait levels, the extreme response style will not contaminate the estimates of raters at different levels.

The extreme answer style allows the IRT models to identify dichotomous and non-scalar questions. For this reason, progress in IRT has been more related to the development of instruments that collect binary data and thus the tradition of multi-item scales observed in Marketing may inhibit increases in the use of IRT method in the field.

Table 3 summarizes the fundamentals of classical and IRT approaches, their advantages and disadvantages, and more adequate applications in individual cases.

Table 3 - Comparison between the IRT and classical approaches of measurement (Factor Analysis and SEM).

ASPECTS	CLASSICAL APPROACHES	ITEM RESPONSE THEORY (IRT)
Fundamentals	They determine individual scores and parameters of items (such as factor loads, average variance extracted and reliability), based on the structure of correlations. The results are not independent of the context where data are collected (collection forms, collection moments and distinct samples).	It calculates parameters to items (adherence to the measured construct) and to raters (trait values) in an independent manner. It usually involves unidimensionality (in simple models) and local independency.
Advantages	Lower complexity. High availability of computer resources. Easy adjustment of multidimensional models. Require smaller samples than IRT models. Generate global reliability indexes.	The items do not have to be cumulative. Greater stability of item parameters in data collected in distinct contexts. Variation in standard error of the item according to the trait level in the rater. It allows for evaluation of reliability by item. With few questions, it is possible to establish the trait value in the individual.
Recommended applications	When there is constant standard error at the trait levels in one item. When we want to have a global indicator of instrument reliability. When we adjust a model with multidimensional traits.	When there is variation in the standard error at the trait levels in one item. When we cannot ensure sample representativeness. To create new measurements and/or refine existing measurements. To obtain reliability by item of the instrument. In the presence of extreme response style.

7.2 Bayesian Estimators

According to Raudenbush & Bryk (2002), classical statistics (not the classical approach of measurement in Marketing, but the classical approach in the field of Statistics) assumes that population parameters are constant and that data used in empiric studies represent probability samples in a universe of possible samples. In the Bayesian approach (based on Bayes' theorem), the idea of probability is not represented by relative frequency in repeated samples, but otherwise by the quantification of the researcher's uncertainty about the unknown parameters that generate the sampled data. In this approach, the parameters have a probability distribution that describes the researcher's uncertainty about their values.

In the classical view, the estimate of a point (and also of a confidence interval) represents a good inference for the parameter value when obtained with a reliable method which is assumed to have adequate theoretical properties. The population parameter is not considered a random variable and, for this reason, it cannot be assigned a probability. In fact, the calculation of the interval that contains the parameter should capture it with some confidence level.

Bayesian statistics, in turn, assume that parameters have a probability distribution and thus we can make inferences based on this principle. A prior distribution describes the researcher's beliefs regarding the parameter before data collection. After data are available, we revise such prior distribution based on data analysis to propose a posterior distribution that combines the evidences from the data with the prior. A point estimate, in this case, can be the central tendency of the posterior distribution (such as its mean or median). An interval estimate in the Bayesian approach can be based on the amplitude of possible values for the parameter, which is the base for the calculation of the posterior probability to have parameter values within the interval.

Prior distributions can incorporate previous knowledge about the parameters, or they can contribute with little information to the posterior distribution when compared to the information from data. This second situation (a prior distribution with little information) refers to priors known as reference priors. Their benefit is that they 'let the data speak'. The application of Bayesian inference is risky when we work with small samples, which, in general, require priors with much information (Congdon, 2006). The risk is in the fact that priors in studies with small samples prevail in the final results of posteriors, reflecting thus the researcher's previous personal judgments. On the other hand, with very large samples, in general, the parameter values estimated with classical approaches and measurements of central tendency of posteriors distributions from the Bayesian approach tend to be the same or very similar (Raudenbush & Bryk, 2002).

The base of contemporaneous Bayesian inference to estimate parameters is the use of Markov Chain Monte Carlo (MCMC) methods, which involve sequential simulations for parameter distributions in long chains (Gamerman & Lopes, 2006). The idea is to summarize the parameters resulting from a MCMC method in the form of expectations, densities and probabilities (Congdon, 2006) obtained by means of simulations using the Monte Carlo principle, not reliable when they are not (approximately) normal or are multimodal.

The original Monte Carlo method involves a set of simulations that are independent of one another. The MCMC methods, in turn, generate pseudorandom simulations by means of Markov chains, in which parameters are considered sequences of random variables. A chain can only be considered a Markov chain if the previous step is relevant to the next step (Rossi, Allenby & McCulloch, 2006), and a simulation using a stable Markov chain converges to a stationary distribution. Thus, a scheme of MCMC simulation converging to stability is established.

There are many questions about how to obtain convergence of MCMC simulation methods. It is usually necessary to establish one initial and short sequence of simulation (burn in period), which will not be used in the final distribution, as the initially simulated parameters may be inadequate; the simulations obtained with MCMC are autocorrelated and thus many of them are required, in order to provide workable results (Rossi *et al*., 2006). In addition, it may take more time to find the posterior density region where the central tendency of the parameter locates, which will depend on the sample size, model complexity and simulation method. If the chains are satisfactorily developed, the autocorrelation will tend to zero as the simulation progresses. Otherwise, little information about the posterior distribution will be provided in every iteration and a larger simulation will be required (Congdon, 2006).

There are several MCMC simulation schemes; the algorithm that serves as the base to all of them is known as Metropolis-Hastings (Congdon, 2006). Another very popular scheme is the Gibbs sampler, an especial case of Metropolis-Hastings algorithm that can simulate marginal distributions in a sequence; and although it generates autocorrelated sequences, it "gets rid of" the initial values of the chain and converges to a stationary distribution.

Especially relevant to the modeling of latent variables is the concept of data augmentation, used to model the likelihood of a model of some nature (such as structural equation modeling); the Gibbs sampler can be used for such purpose. The concept of data augmentation refers to adding unavailable information (such as the estimation of latent variables) to the set of data by modeling. Rossi, Allenby & McCulloch (2006) demonstrate that a number of models can be constructed with data augmentation when we cannot observe variables directly. To see more about MCMC simulation algorithms, we recommend reading the books of Gamerman & Lopes (2006) and Rossi, Allenby & McCulloch (2006).

Particularly regarding the use of a confirmatory factor analysis in construct validation, the Bayesian approach has some advantages in relation to the classical statistical approach. First, the researchers who prefer to use the Bayesian inference claim that it allows to use smaller samples than in the classical approach (Rossi, Allenby & McCulloch, 2006); however, this argument is only true when we have priors with much information, which, as mentioned above, is risky. To mitigate this risk, we suggest that researchers conduct an extensive literature revision and analysis of empiric results to define prior distributions that they will use in their models.

Another advantage is that the use of Bayesian estimators does not require the violation of the distributions of the variables employed. Most items of scales in Marketing are collected with ordinal variables (degrees of agreement, for instance), but treated through classical statistical models that assume that the collected data are normally distributed, just like maximum-likelihood estimation, frequently used in confirmatory factor analysis and structural equations modeling. The normal distribution of an ordinal variable is very improbable, or even impossible, considering that the normal distribution is for continuous quantitative variables. As it does not presume a normal distribution of data, the Bayesian inference is better used in ordinal variable modeling (Byrne, 2001).

The classical approach sometimes relies on asymptotic approximations to provide a density function of probability to the set of estimators. Even considering asymptotic approximations do not presume data normality, they do not perform sufficiently in non-linear models, unlike Bayesian estimation (see Zellner & Rossi, 1984). In addition, modeling with asymptotic approximations requires very large samples, a clear disadvantage in relation to Bayesian modeling. Finally, the Bayesian models are less sensitive to the presence of outliers, as the distribution of parameters is mostly based on the majority of the sample and less on extreme cases (Hahn & Doh, 2006).

For considering the existence of a possible distribution of parameters in the population, and not the existence of a constant population parameter, some authors consider Bayesian inference as the most adequate (or the only) method to adjust models in Marketing (see Rossi, Allenby & McCulloch, 2006; and Park & Kim, 2013). They claim that it is possible to model behaviors and attitudes of every individual based on their particular characteristics, instead of estimating an average parameter for the whole population (a limitation of classical statistical models). As in Marketing it is relevant to understand the agents in a personalized manner, this property of Bayesian models has driven to the use of this type of inference in this field. Such benefit of the Bayesian methods is similar of one benefit brought by the IRT models, which, due to the modeling of raters' parameters, can also be considered Bayesian in their nature. However, Bayesian estimation in factor analysis and structural equations, unlike IRT, does not separate item parameters from rater parameters and it is based on the association between data (just like the classical approach) and, therefore, depends on the sample characteristics. Indeed, when we use reference priors, posterior distributions of parameters fit well the sample which, therefore, should be representative of the population.

Bayesian models have been increasingly used in several fields due to its intuitive nature and advantages in relation to classical inference. This development has been driven by increasing software that can provide Bayesian estimation with user-friendly interfaces, as well as hardware developments that can process simulations of very large sequences (thousands). One example of application of this type is the algorithm present in AMOS structural equation package. However, these tools tend to provide little flexibility to researchers in terms of MCMC simulation selection or extraction of individualized results for each rater, which would be one of the main benefits of choosing a Bayesian model. There are more flexible tools, such as R software, which require, however, advanced statistical knowledge and programming skills, usually not very common among the skills that researchers in Marketing have.

The use of classical inference by means of maximum likelihood in the construction of scales is adequate when we observe data normality and the absence of outliers, but this is not probable, and such a scenario favors the use of Bayesian models. The MCMC models are also preferable when we have small samples but, but in this case, we need priors to be informative, rigorously determined by coherent theoretical formulations and the analysis of results from previous studies; otherwise, parameters will be very dependent on idiosyncrasies of the researcher that formulates the prior.

The use of continuous variables to rate indicators allows for maximum-likelihood and Bayesian estimation, while the use of ordinal and/or discrete is more suitable to the Bayesian estimation. Table 4 summarizes the fundamentals of classical inference and Bayesian inference in factor analysis and structural equation modeling, their advantages and disadvantages, and the more adequate applications in individual cases.

Table 4 - Comparison between classical estimation and Bayesian estimation in Factor Analysis and Structural Equation Modeling.

ASPECTS	CLASSICAL ESTIMATION	BAYESIAN ESTIMATION
Fundamentals	Assumes the existence of a fixed parameter in the population and calculates the confidence interval using a classical approach (in general, by means of maximum likelihood).	Assumes the existence of a distribution of parameters in the population and estimates it based on a prior formulation, to be improved to a distribution that considers the data collected, i.e., a posterior distribution. Estimation is made by using Markov Chain Monte Carlo (MCMC) simulations.
Advantages	More common in statistical software of simple execution.	It does not provide data distribution normality; it is not limited to modeling with continuous quantitative variables; it is not very sensitive to discrepant data. It can estimates individual parameters to raters, instead of one average parameter to the population.
Recommended applications	Large samples, with normally distributed data, absence of outliers, continuous quantitative variables.	Small samples, with variables of different nature (e.g., qualitative and discrete variables, besides quantitative variables), presence of outliers.

7.3 Modeling by partial least squares (PLS)

The use of confirmatory factor analysis and structural equation modeling in Marketing, as mentioned before, has been frequent with the use of maximum-likelihood estimation. One viable option that has become popular in empiric studies in the area, with strong influence of the field of Information Systems, is the use Partial Least Squares (PLS) modeling. Although covariance-based models (such as maximum-likelihood estimation) are more frequently used by researchers in Marketing, the PLS models are also applications of structural equation, but based on variance.

The most relevant difference between covariance-based models and variance-based models is that the first provide global adjustment indexes by comparing covariance matrices (or correlation matrices, in standardized models) estimated by the model to those actually observed in the collected data and chi-square-based tests are conducted to evaluate the differences. The general adjustment of the model is accessed by indicators based on chi-square statistics and tests that evaluate the significance of errors derived from the difference between what is observed and what is estimated. In variance-based models, there are no global statistics adjustments and the model is evaluated by means of the significance of the relations proposed among the variables (also available in covariance-based models) and by the total variability of the variables of interest that the model can explain (R^2).

As it does not depend on a covariance structure to adjust a model, the PLS method has been used in studies that use formative indicators, as they do not require high correlations between the indicators used to measure the same construct. This has been considered by many researchers that conduct studies with formative indicators a good reason to use the method, but it has also been the focus of discussions on its fragilities (see Diamantopoulos, 2011). For not having global adjustment indicators, PLS models are limited to analyzing whether the proposed relations make sense individually, but they do not allow us to analyze whether the mode is plausible as a whole. For this reason, the literature usually recommends PLS methods to exploratory studies which do not count on advanced theoretical developments (see Hair, Ringle & Sarstedt, 2011; Henseler, Ringle & Sinkovics, 2009; Marcoulides & Saunders, 2006; and Ringle, Sarstedt & Straub, 2012). However, the adequacy of this situation is questionable in procedures for nomological validity, in which the theoretical relations between constructs and variables should be well developed by the researcher, and that is usually the case favoring the use of structural equations.

We should note that covariance-based models (except for factor analysis; and confirmatory factor analysis is a specific case of structural equation modeling) are not necessarily models for reflective indicators. However, in practical terms, they end up as such; because they usually count on the covariance of items used to measure a construct, they are often not 'identified' (and so do not converge to an adjustment) when such a covariance is not stipulated or when it is not high enough to make the model ‘run’. For this reason, even if covariance-based models are more recommended to nomological validity with measurements of different nature (formative or reflective, for having global adjustment indicators), it will sometimes be difficult to converge them, since they involve greater complexity to be adjusted (Diamantopoulos, 2011).

One problem of PLS models is the fact that they do not allow error estimation to formative indicators, while variance-based models estimate such kind of errors; we understand that it is not reasonable not to estimate measurement errors.

Other applications of the PLS method found in the literature are cases without large samples, as fewer parameters are estimated when compared to covariance-based models, and thus, we spare degrees of freedom (although we know that, in classical estimation, confidence intervals are greater for small samples, that is, less precise). In addition, variance-based models do not require normality of data distribution. The use of PLS models has increased due to software availability with user-friendly interfaces, such as SmartPLS and PLS-Graph.

Table 5 summarizes the fundamentals of structural equation modeling based on covariance (particularly, maximum-likelihood estimation) and on variance (particularly, estimation by means of Partial Least Squares - PLS), their advantages and disadvantages, and more adequate applications in particular cases.

Table 5 - Comparison between structural equation modeling by maximum-likelihood estimation and Partial Least Squares (PLS).

ASPECTS	MAXIMUM-LIKELIHOOD ESTIMATION	PLS
Fundamentals	Obtains global adjustment indexes when comparing covariance matrices estimated by the model to those actually observed in the collected data and by conducting chi-square-based tests to evaluate the differences. It also accesses the significance of relations proposed among the variables of interest.	Accesses the significance of relations proposed among variables and total variability of variables of interest that the model can explain (R^2).
Advantages	Calculates global adjustment indexes of the model, some with significance tests. Estimates errors to formative measurements, if any.	Does not presume data distribution to be normal. Requires smaller samples.
Recommended applications	Large samples, with normally distributed data. Presence of reflective indicators only.	Smaller samples, data without normal distribution. Studies that use formative indicators.

8 FINAL CONSIDERATIONS

Using a recommendation of Pedhazur & Schmelkin (1991), to be meaningful, any research activity, including scientific article reading, should be based on critical considerations. We recommend that students from scientific fields in masters and doctoral degrees programs and researchers should keep critical thinking on their methodological choices, especially when involving measurement.

In fact, there is no sense in using sophisticated statistical modeling if the database for such analyses present numbers that do not properly reflect the phenomena they should. Keeping that in mind, we’ve developed in this article a broad revision of the historical evolution, the current scenario and trends of the construct measurement issues in Marketing.

In our view, the academic and professional development of Marketing depends on the development of researches to improve knowledge in the field. But research in Marketing is, in turn, dependent on the level of methodological development, which involves issues of measurement, design and data analysis. Definitely, there may be no development of solid research in Marketing without careful focus on measurement of variables and theoretical constructs. Aligned with the perception of Lee & Hooley (2005), we recommend that researchers in Marketing should dedicate as much time as necessary to make solid measurement models; only later it will make sense to elaborate advanced studies to test assumptions among constructs.

Our article has provided a broad discussion on this subject. We attempted to include every core theme about the subject, which makes us believe that, in an academic perspective, our study is useful to both inexperienced or experienced researchers that seek updated information and an overview of the subject.

In addition, we understand this article brings potential contribution to the field of education in Marketing, especially to the area of Marketing

Research, taught in graduate courses or methodological content disciplines in postgraduate programs. Therefore, this article can be used as a component of a more general discipline, as well as an introductory text of a more specific discipline about measurement (we already have examples of disciplines of this nature in postgraduate programs in Brazil, for example at EAESP/FGV (São Paulo), ESPM (São Paulo), FUMEC (Minas Gerais) and UFPB (Paraíba).

This article shows that we have advanced in theoretical terms, with increasing contribution from Marketing researchers to the theme of measurement (unlike the past scenario, when the field of Marketing depended on developments conducted in the fields of Psychology and Education). Our challenge to Brazilian researchers is to keep studying the subject, expanding boundaries and seeking to develop the analysis of this theme. The content of this article also shows that we still have to progress and that the challenges are very encouraging. Now, our demand lies in the development of other studies and applications to further improve the knowledge we produce in Marketing.

REFERENCES

American Psychological Association. (1985). Standards for educational and psychological tests. Washington, DC: Author.

Andrade, D. F., Tavares, H. R., & Valle, R. C (2000). Teoria da resposta ao item: conceitos e aplicações. *14º Simpósio Nacional de Probabilidade e Estatística – SINAPE*. São Paulo: Associação Brasileira de Estatística.

Aranha, F., & Zambaldi, F. (2008). *Análise fatorial em administração*. Sao Paulo: Cengage Learning.

Bagozzi, R. P., & Edwards, J. R. (1998). A general approach for representing constructs in organizational research. *Organizational Research Methods*, *1*(1), 45-87.

Barboza, S. I. S., Carvalho, D. L. T., Soares Neto, J. B. & Costa, F. J. (2013). Variações de Mensuração pela Escala de Verificação: uma análise com escalas de 5, 7 e 11 pontos. *Teoria e Prática em Administração*, *3*(2), 99-120.

Belk, R. W. (1985). Materialism: trait aspects of living in the material world. *Journal of Consumer Research*, *12*(3), 265-280.

Buchbinder, F., Goldszmidt, R., & Parente, R. (2012). Item Response Theory and Construct Measurement in Emerging Markets. *Research Methodology in Strategy and Management*, *7*, 73-100.

Bussab, W. O., & Morettin, P. (2007). Estatística Básica. São Paulo: Saraiva.

Byrne, B. (2001). *Structural Equation Modeling with Amos: Basic Concepts. Applications. and Programming*. Mahwha, New Jersey: Lawrence Erlbaum.

Church, A. (2010). Measurement issues in cross-cultural research. In G. Walford, E. Tucker, & M. Viswanathan (Eds.), *The Sage Handbook of Measurement* (pp. 151-176). London, UK: Sage Publications.

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

Congdon, P. (2006). *Bayesian Models for Categorical Data*. Chichester, England: John Wiley &Sons, Ltd.

Costa, F. J. (2011). Mensuração e Desenvolvimento de Escalas. Rio de Janeiro: Editora Ciência Moderna.

Cronbach, L. J. (1951). Coefficient Alpha and the Internal Structure of Tests. *Psychometrika*, *16*(3), 297-334.

de Jong, M. G., Steenkamp, J.-B. E. M., Fox, J.-P., & Baumgartner, H. (2008). Using Item Response Theory to Measure Extreme Response Style in Marketing Research: A Global Investigation. *Journal of Marketing Research*, *45*(1), 104-115.

Devellis, R. F. (1991). *Scale development*: theory and applications. Newbury Park, CA: SAGE Publications.

Diamantopoulos, A. (2011). Incorporating formative measures into covariance-based structural equation models. *Mis Quarterly*, *35*(2), 335-358.

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

Edwards, J. R. (2011). The fallacy of formative measurement. *Organizational Research Methods, 14*(2), 370-388.

Faraway, J. J. (2006). *Extending linear models with R*. Boca Raton, FL: Chapman & Hall/CRC.

Fornell, C., & Larcker, D. F. (1981). Structural equation models with unobservable variables and measurement error: Algebra and statistics. *Journal of marketing research, 18*(8), 382-388.

Gamerman, D., & Lopes, H. (2006). *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. Boca Raton, FL: Chapman & Hall/CRC.

Gerbing, D. W., & Anderson, J. (1988). An Updated Paradigm for Scale Development Incorporating Unidimensionality and Its Assessment. *Journal of Marketing Research*, *25*, 186-192.

Gonçalves, H. M. M. (2013). Multi-group invariance in a third-order factorial model: attribute satisfaction measurement. *Journal of Business Research*, *66*, 1292-1297.

Hahn, E. D., & Doh, J. P. (2006). Using Bayesian methods in strategy research: an extension of Hansen et al. *Strategic Management Journal*, *27*(8), 783-798.

Haig, B. D. (2005). An abductive theory of scientific method. *Psychological Methods*, *10*(4), 371-388.

Hair, J. F., Ringle, C. M., & Sarstedt, M. (2011). PLS-SEM: Indeed a silver bullet. *The Journal of Marketing Theory and Practice*, *19*(2), 139-152.

Hao, L. & Naiman, D. Q. (2007). *Quantile regression*. Thousand Oaks: Sage Publications.

Henseler, J., Ringle, C. M., & Sinkovics, R. R. (2009). The use of partial least squares path modeling in international marketing. *Advances in international marketing*, *20*(1), 277-319.

Hodge, D. R. & Gillespie, D. F. (2007). Phrase completion scales: a better measurement approach than Likert scales? *Journal of Social Service Research*, *33*(4), 1-12.

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

Kamakura, W. A., & Mazzon, J. A. (2013). Socioeconomic status and consumption in an emerging economy. *International Journal of Research in Marketing, 30*(1), 4-18.

Kloke, J. D., & Mckean, J. W. (2012). Rfit : Rank-based estimation for linear models. *The R Journal*, *4*(2), 57–64.

Lee, C. E. (1965). Measurement and the development of science and marketing. *Journal of Marketing Research*, *2*(1), 20-25.

Lee, N., & Hooley, G. (2005). The evolution of "classical mythology" within marketing measure development. *European Journal of Marketing*, *39*(3), 365-385.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives in Psychology*, *140*, 1-55.

Lucian, R. (2012). *Mensuração de atitudes*: a proposição de um protocolo para elaboração de escalas. Tese (Doutorando em Administração). Programa de Pós-Graduação em Administração da Universidade Federal de Pernambuco – PROPAD-UFPE.

Marcoulides, G. A., & Saunders, C. (2006). Editor's comments: PLS: a silver bullet?. *MIS quarterly*, *30*(2), iii-ix.

Mari, L. (2005). The problem of foundations of measurement. *Measurement*, *38*(4), 259-266.

McDonald, R. P. (2000). A basis for multidimensional item response theory. *Applied Psychological Measurement*, *24*(2), 99–114.

Meade, A. W., & Lautenschlager, G. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods*, *7*(4), 361-388.

Milfont, T. L., & Fischer, R. (2010). Testing measurement invariance across groups: Applications in cross-cultural research. *International Journal of Psychological Research*, *3*(1), 111-121.

Netemeyer, R. G., Bearden, W. O., & Sharma, S. (2003). *Scaling procedures*: issues and applications. Thousand Oaks: Sage.

Nunnaly, J. (1978). *Psychometric Theory*. New York: McGraw-Hill Book Company.

Park, H. J., & Kim, S. H. (2013). A Bayesian network approach to examining key success factors of mobile games. *Journal of Business Research*, *66*(9), 1353-1359.

Pedhazur, E., & Schmelkin, L. P. (1991). *Measurement, design and analysis*: an integrated approach. Hillsdale: Lawrence Erlbaum Associates Inc. Publishers, 1991.

Pereira, B. B. (1997). Estatística: a tecnologia da ciência. *Boletim da Associação Brasileira de Estatística*, ano XIII, n. 37, 2º quadrimestre, 27-35.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: applications and data analysis methods*. (2nd ed.). Thousand Oaks, CA: Sage Publications, Inc.

Reardon, J., & Miller, C. (2012). The effect of response scale type on cross-cultural construct measures: an empirical example using Hall's concept of context. *International Marketing Review*, *29*(1), 24-53.

Richins, M. L., & Dawson, S. (1992). A Consumer Values Orientation for Materialism and Its Measurement: Scale Development and Validation. *Journal of Consumer Research*, *19*(3), 303-316.

Ringle, C. M., Sarstedt, M., & Straub, D. W. (2012). Editor's comments: a critical look at the use of PLS-SEM in MIS quarterly. *MIS quarterly*, *36*(1), iii-xiv.

Rossi, P., Allenby, G., & McCulloch, R. (2006). *Bayesian statistics and marketing*.

Chichester, England: John Wiley and Sons, Ltd.

Rossiter, J. R. (2002). The COARSE procedure for scale development in marketing. *International Journal of Research in Marketing*, *19*(4), 305-335.

Rossiter, J. R. (2011) *Measurement for the Social Sciences*: the COARSE method and why it must replace psychometrics. New York: Springer.

Salzberger, T., & Koller, M. (2013). Towards a new paradigm of measurement in marketing. *Journal of Business Research, 66*(9), 1307-1317.

Samartini, A. L. S. (2006). *Modelos com variáveis latentes aplicados à mensuração de importância de atributos*. Doctoral thesis, Escola de Administração de Empresas de São Paulo da Fundação Getulio Vargas (FGV/EAESP), Sao Paulo, Brazil.

Scherbaum, C., Finlinson, S., Barden, K., & Tamanini, K. (2006). Applications of item response theory to measurement issues in leadership research. *The Leadership Quarterly*, *17*(4), 366–386.

Sheather, S. J. (2009) *A modern approach to regression with R*. New York: Springer.

Steenkamp, J. -B. E. (2005). Moving out of the US silo: A call to arms for conducting international marketing research. *Journal of Marketing*, *69*(4), 6-8.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 677–680.

Stewart, D. W. (1981). The application and misapplication of factor analysis in marketing research. *Journal of Marketing Research*, *18*(2), 51-62.

Urbina, S. (2004). *Essentials of psychological testing*. New Jersey: John Wiley & Sons, Inc..

Van de Vijver, F. J., & Leung, K. (1997). Methods and data analysis for cross cultural research. Thousand Oaks, CA: SAGE.

Wong, N., Rindfleisch, A., & Burroughs, J. E. (2003). Do reverse-worded items confound measures in cross-cultural consumer research? The case of the material values scale. *Journal of Consumer Research*, 30(1), 72–91.

Yi, Y., & Gong, T. (2013). Customer value co-creation behavior: scale development and validation. *Journal of Business Research*, *66*, 1279-1284.

Zellner, A., & Rossi, P. E. (1984). Bayesian analysis of dichotomous quantal response models. *Journal of Econometrics*, *25*(3), 365-393.