Original Article

D-test: A New Test for Analyzing Scale Invariance Using Symbolic Dynamics and Symbolic Entropy

Jose A. Martínez[1] and Manuel Ruiz Marín[2]

[1]Department of Management and Marketing, Technical University of Cartagena, Spain, [2]Department of Quantitative Methods and Computing, Technical University of Cartagena, Spain

Abstract. The aim of this study is to improve measurement in marketing research by constructing a new, simple, nonparametric, consistent, and powerful test to study scale invariance. The test is called *D*-test. *D*-test is constructed using symbolic dynamics and symbolic entropy as a measure of the difference between the response patterns which comes from two measurement scales. We also give a standard asymptotic distribution of our statistic. Given that the test is based on entropy measures, it avoids smoothed nonparametric estimation. We applied *D*-test to a real marketing research to study if scale invariance holds when measuring service quality in a sports service. We considered a free-scale as a reference scale and then we compared it with three widely used rating scales: Likert-type scale from 1 to 5 and from 1 to 7, and semantic-differential scale from −3 to +3. Scale invariance holds for the two latter scales. This test overcomes the shortcomings of other procedures for analyzing scale invariance; and it provides researchers a tool to decide the appropriate rating scale to study specific marketing problems, and how the results of prior studies can be questioned.

Keywords: measurement, rating scales, scale invariance, *D*-test, symbolic dynamics, symbolic entropy

Rating scales are probably the most popular measurement method in marketing research. In 2002, an estimated US$ 6.8 billion was spent on conventional marketing research tools in the United States alone (Mast & Zaltman, 2005). In 2007, this number grew up to US$ 12 billion (Lindstrom, 2008). A great portion of this enormous amount of money was invested in quantitative techniques using standard questionnaires administered in written or telephone interviewing, being rating scales the preferred method for measuring verbal self-reports: attitudes, opinions, and intentions.

Since the first decades of the past century, social science researchers have been searching for an ideal measurement method that minimizes bias and maximizes utility.[1] Rating scales have been the most used measurement method since Likert (1932) introduced his scale to measure attitudes. However, during the last several decades, researchers have discussed numerous concerns related to how to implement this method. Several examples of these debates are: (1) the proper number of response alternatives in the rating scale; (2) the choice of verbal labels to identify the different alternatives; (3) reliability and validity of the scales with different anchors; (4) the effect of incorporating a neutral point; and (5) consumer's preferences among the different response formats. The scholarly discussion of the issues created several alternatives such as semantic-differential scales and other categorical measurement scales with limited response alternatives. There are excellent references in the literature that review the debate, such as Cox (1980) or Hofmans, Theuns, and Mairesse (2007). These studies also demonstrated numerous contradictions that have been derived from empirical research.

In marketing research the most used rating scales are 5- and 7-point Likert-type scales. The studies of Cox (1980) or Lissitz and Green (1975) are frequently referenced by researchers in order to support the choice of the mentioned Likert-type scales. Likewise, the use of semantic-differential scales is also very widespread, because of the contribution of prospect theory (Kahneman & Tversky, 1979) and the distinction between positive and negative emotions. However, researchers need to know if a specific rating scale is statistically less or equally useful than other rating scale. In other words, researchers need to analyze if scale invariance holds, that is if the measure of the variable of interest does not change when the length of the scale is multiplied by a constant

[1] We will use the term "utility" instead of "validity." We agree with the viewpoint of Borsboom, Mellenbergh, and van Heerden (2004) regarding the concept of validity and its distinction from the concept of utility. Therefore, a test is valid for measuring an attribute if (a) the attribute exists and (b) variations in the attribute causally produce variation in the measurement outcomes. However, a valid test may be more useful than another valid test for measuring the same attribute if it is a better representation of the measured attribute. In words, a span and a meter are two valid forms to measure distance, but the meter is more useful than the span. There are no degrees of validity but degrees of utility.

DOI: 10.1027/1614-2241/a000026

factor, for example, when a scale used to measure a construct in a specific study is a 10-point scale and the proposed scale for measuring the same construct in other study is a 5-point scale. We consider deviations from scale invariance as bias, and the objective of this research is to develop a statistical test to make this bias quantifiable.

In this article, we construct a new, simple, nonparametric, consistent, and powerful test to compare the response pattern that comes from a specific rating scale (a free-scale), with the response pattern that comes from other rating scales. We also applied the proposed test to a real marketing research, comparing the free-scale with three widely used rating scales: Likert-type scale from 1 to 5 and from 1 to 7, and semantic-differential scale from −3 to +3. The test is called *D*-test, and it is constructed using symbolic dynamics and symbolic entropy as a measure of the difference between the response patterns. Symbolic entropy was first introduced in the context of time series analysis in Matilla and Ruiz (2008). We also give a standard asymptotic distribution of our statistic. Given that the test is based on entropy measures, it avoids smoothed nonparametric estimation.

The relevancy of the proposed *D*-test is twofold: firstly, this test overcomes the shortcomings of other procedures for analyzing scale invariance; and secondly, it provides researchers a tool to decide the proper rating scale for studying specific marketing problems, and how the results of prior studies can be questioned. This latter fact is especially important in meta-analysis.

The rest of this paper is organized as follows. In the following section, the notion of scale invariance is defined; this includes a review of some procedures for testing this assumption. The paper then details the construction of the *D*-test followed by its application to a real marketing research (service quality in a sports center) in order to illustrate our approach. The paper concludes with a brief discussion about the implications for marketing research.

Scale Invariance

Scale invariance is a characteristic of objects that does not change if the length of the scale is multiplied by a constant factor. For example, given the polynomial function $f(x) = ax^k$, where a and k are constants, $f(cx) = c^k ax^k = c^k f(x)$, where c is a constant, that is, by scaling the argument of the function by a constant factor c, the function is rescaled by a constant factor c^k.

Steven's (1951) work showed that the response to a stimulus $f(x)$ is a power function of the evoked sensation x, so $f(x) = ax^k$. As $k = 1$ for line length production and numeric estimation (Hofmans, Theuns, Baekelandt, et al., 2007), then $\ln(f(x)) = \ln(a) + \ln(x)$. If the response to the stimulus is given in a free-scale, it would be plausible to assume that $a = 1$, then the relationship between response and evoked sensation would be identical. If the same phenomenon is measured using other rating scale, then $\ln(f(cx)) = \ln(c) + \ln(a) + \ln(x)$. Consequently, we obtain the rescaled factor by finding c.

As it can be intuited, transformation from a 10-point scale to a 5-point scale implies categorization bias (see Cox (1980) for an illustration). The relevant question is whether this potential bias is statistically significant.

In the following subsections we briefly depict some procedures to analyze scale invariance.

Direct Comparison

The first and simpler option is to make a direct comparison between means and variances of the two scales (A and B) normalized to the unit interval. If both statistics do not differ, this would favor the hypothesis that scale invariance holds.

The shortcoming of this approach is that means and variances could be statistically equal, but the response pattern of each individual could not be invariant.

In the case of the mean, we may assume that if both means are equal, the expected value of all individual deviations from scale invariance is zero. Consequently, we would consider these deviations as a categorization error with zero mean. The variance of this error would distort the variance of scale B, so that the hypothesis of equality of variances would be rejected with more assiduity. However, we strongly believe that assuming that this categorization error is a random variable with zero mean is not a very realistic assumption, since this would imply that individuals would always be capable of mapping mentally every value of the scale A to the proper category of the scale B (the nearest integer). In addition, this assumption would require the distribution of responses in the scale A to be uniformly distributed or perfectly symmetrical, in order to balance positive and negative categorization errors. This latter fact is certainly difficult to assume in practice.

Therefore, by using this method we do not know if the hypothesis of scale invariance holds simply by chance.

Structural Equation Modeling

A more sophisticated method based on the covariances of several measurement scales can be used. Structural equation modeling is a widely employed methodology in social sciences that consider latent variables and observable indicators (individual's responses).

We may build the simplest reflective model relating a nonobservable latent variable (what we want to measure) with the observable indicator (the item measured on a rating scale). The equation is the following: $y = \lambda\varepsilon + \delta$, being y the value of the observable indicator, ε the value of the latent variable, λ the regression coefficient that relates both values, and δ an error term. The analogy of this equation with the equations defined for the scale invariance is quickly viewed. In this case, it is assumed that the error term follows a normal distribution with zero mean. Therefore, the error term does not distort the mean value of the responses, but increases their variance. Note that one of the strengths of structural equation modeling is the possibility to consider measurement error in the observable indicators.

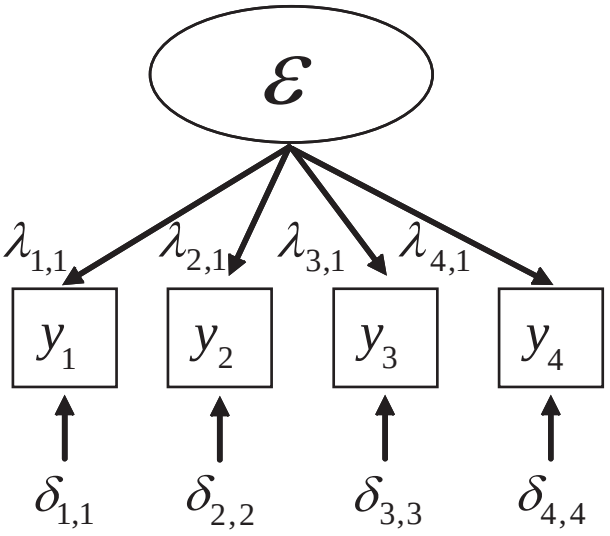

Figure 1. Measurement model for testing scale invariance.

The previous equation may be expressed in terms of covariance by the following expression: $Var(y) = \lambda^2 Var(\varepsilon) + Var(\delta)$, assuming independence on the right-hand side of the equation. Scale invariance can be studied with four scales, the free-scale: y_1, and other three distinct rating scales: y_2, y_3, and y_4. Figure 1 illustrates the model.

The main advantage of structural equation modeling is that it allows the analysis of different degrees of scale invariance. We assume that the relationship between the response and the evoked sensation is identical (so $\lambda_{1,1} = 1$). As all data are normalized to a unit interval, if scale invariance holds, the remaining lambda parameters should also be equal to 1. In structural equation modeling terminology this situation is called *tau-equivalent indicators model.* Moreover, if we also assume that measurement error is different from zero ($Var(\delta_{1,1}) > 0$), a more restricted type of scale invariance would occur if the remaining theta parameters are equal to $Var(\delta_{1,1})$. We need to fix $Var(\delta_{1,1})$ to a specific value (see Hayduk, 1996). This situation is known as *parallel indicators model.* A detailed explanation of these models can be found in Jöreskog and Sörbom (2001).

Model illustrated in Figure 1 is identified, so researchers can test different models considering distinct restrictions in the estimated parameters in order to test scale invariance (tau-equivalent and parallel models). All these models can be compared using the chi-square difference test in a nested model sequence (see Yuan & Bentler, 2004).

The shortcomings of this way of testing scale invariance are associated to the general shortcomings of the structural equation modeling methodology (e.g., Tomarken & Waller, 2005). In general, it is desirable to count with continuous data in order to apply maximum likelihood estimation.[2] However a minimum sample size is needed in order to compute the asymptotic covariance matrix and to avoid bias in the chi-square statistic (Jackson, 2003). When data are not continuous, asymptotic free estimation is necessary, and requirements about sample size are much more restrictive. In this case, the input is not a covariance matrix but a polyserial or polychoric correlation matrix, thus interpretation of measurement error as a percentage of variance of the latent variable is more complicated. This latter fact makes it difficult to fix the parameters of the free-scale for testing tau-equivalent and parallel models.

Furthermore, this type of methodology requires that models have to be identified, so a minimum number of indicators is needed for each specific model susceptible to be tested. For example, the model illustrated in Figure 1 is not identified with two indicators; consequently, a simple comparison between two scales would not be feasible.

The procedure that we introduce in this paper overcomes some of the limitations of structural equation modeling, being less restrictive regarding the nature of data.

D-test: Preliminaries and Notation

In this section we give some definitions and introduce the basic notation.

Let P be the population to be studied. Let N be the number of individuals in the population P. Suppose that every individual $e \in P$ is asked to evaluate some item. In order to do so we let the individual to chose his/her own scale, namely scale A, and afterwards we ask this individual to evaluate the same item but with another scale, namely B. Assume that scale A is within the interval $[a_1,a_2]$ and scale B is within the interval $[b_1,b_2]$. Denote the value of the evaluation of individual e in the scale A by x_e and in the scale B by y_e. In order to normalize both scales to the unit interval [0,1] we rescale as follows:

$$\tilde{x}_e = \frac{x_e - a_1}{a_2 - a_1} \quad \text{and} \quad \tilde{y}_e = \frac{y_e - b_1}{b_2 - b_1}.$$

Now divide the unit interval [0,1] in k subintervals, $I_1, I_2, \ldots, I_k$, of the same length, that is

$$I_i = \left[\frac{i-1}{k}, \frac{i}{k}\right]$$

for every $i = 1, 2, \ldots, k$. Denote by $S = \{I_1, I_2, \ldots, I_k\}$. Then every valuation $\tilde{x}_e$ and $\tilde{y}_e$ belongs to at most two of the intervals in S. Then, we can associate to every individual $e \in P$ the value (I_j, A) if $\tilde{x}_e \in I_j$ or the value (I_j, B) if $\tilde{y}_e \in I_j$. In order to ensure the uniqueness of this association we impose the following condition, if $\tilde{x}_e \in I_j \cap I_{j+1}$

[2] Recall that data coming from rating scales should be considered as categorical data. There exists a debate about this topic. Our view in this paper is the following: If individuals prefer to give their responses in a 10-point scale, but they have to respond in a 5- or 7-point scale, then data coming from these responses will be categorical. Therefore, maximum likelihood estimation in structural equation modeling should not be used.

(resp. $\tilde{y}_e \in I_j \cap I_{j+1}$), that is $\tilde{x}_e = \frac{j}{k}$ (resp. $\tilde{y}_e = \frac{j}{k}$), then the associated value is (I_j, A) (resp. (I_j, B)).

We will call an element in $S \times \{A, B\}$ a symbol. Therefore we can define the following map

$$f : P \times \{A, B\} \to S \times \{A, B\}$$

defined by $f(e, t) = (\alpha, t)$ for $\alpha \in S$ and $t \in \{A, B\}$, that is, the map f associates to each individual $e \in P$ the valuation of the item and whether e used scale A or B . We will call f a *symbolization map.* In this case we will say that individual e is of (α, t)-type.

Denote by

$$n_\alpha = \sharp\{(e, A) \in P \times \{A, B\} | f(e, A) = (\alpha, A)\},$$

and

$$m_\alpha = \sharp\{(e, B) \in P \times \{A, B\} | f(e, B) = (\alpha, B)\},$$

that is, the cardinal of the subsets of $P \times \{A, B\}$ formed by all the individuals of (α, A)-type and (α, B)-type, respectively. Therefore $n_\alpha + m_\alpha$ is the number of individuals of α-type.

Also, under the conditions above, one could easily compute the relative frequency of a symbol $(\alpha, t) \in S \times \{A, B\}$ by:

$$p_\alpha = \frac{\sharp\{(e, A) \in P \times \{A, B\} |\ e \text{ is of } (\alpha, A) - \text{type}\}}{2N} \tag{1}$$

and

$$q_\alpha = \frac{\sharp\{(e, B) \in P \times \{A, B\} |\ e \text{ is of } (\alpha, B) - \text{type}\}}{2N}. \tag{2}$$

Hence the total frequency of a symbol α is $s_\alpha = p_\alpha + q_\alpha$.

Now under this setting we can define the *symbolic entropy* of a valuation. This entropy is defined as the Shanon's entropy of the k distinct symbols as follows:

$$h(S) = -\sum_{\alpha \in S} s_\alpha \ln(s_\alpha). \tag{3}$$

Similarly we have the symbolic entropy for the valuation with scale A and with scale B

$$h(S, A) = -\sum_{\alpha \in S} p_\alpha \ln(p_\alpha), \tag{4}$$

and

$$h(S, B) = -\sum_{\alpha \in S} q_\alpha \ln(q_\alpha), \tag{5}$$

respectively.

Symbolic entropy, $h(S)$, is the information contained in comparing the k symbols in S among all the individuals in P.

Construction of the Test

In this section we construct a test to detect when valuation of the evaluated item is affected by the chosen scale with all the machinery defined in *D*-test. Preliminaries and Notation section. In order to construct the test, which is the aim of this paper, we consider the following null hypothesis:

$$H_0 :\ \text{the scale does not affect the valuation of the item (scale invariance holds)}, \tag{6}$$

that is,

$$H_0 : q_\alpha = p_\alpha \text{ for all } \alpha \in S \tag{7}$$

against any other alternative.

Now for a symbol $(\alpha, t) \in S \times \{A, B\}$ and an individual $e \in P$ we define the random variable $Z_{(\alpha,t)e}$ as follows:

$$Z_{(\alpha,t)e} = \begin{cases} 1 & \text{if } f(e, t) = (\alpha, t) \\ 0 & \text{otherwise,} \end{cases}$$

that is, we have that $Z_{(\alpha,t)e} = 1$ if and only if e is of (α, t)-type, $Z_{(\alpha,t)e} = 0$ otherwise.

Then $Z_{(\alpha,t)e}$ is a Bernoulli variable with probability of "success" either p_α if $t = A$ or q_α if $t = B$, where "success" means that e is of (α, t) -type.

Then we are interested in knowing how many e's are of (α, t)-type for all symbols $(\alpha, t) \in S \times \{A, B\}$. In order to answer the question we construct the following variable

$$Y_{(\alpha,t)} = \sum_{e \in P} Z_{(\alpha,t)e}. \tag{9}$$

The variable $Y_{(\alpha,t)}$ can take the values {0, 1, 2, …, N}. Then it follows that the variable $Y_{(\alpha,t)}$ is the binomial random variable

$$Y_{(\alpha,A)} \approx B(N, p_\alpha) \tag{10}$$

or

$$Y_{(\alpha,B)} \approx B(N, q_\alpha). \tag{11}$$

Then the joint probability density function of the $2k$ variables

$$P(Y_{(I_1,A)} = a_1, \ldots, Y_{(I_k,A)} = a_k, Y_{(I_1,B)} = b_1, \ldots, Y_{(I_k,B)} = b_k)$$

is

$$\frac{(a_1 + \ldots + a_k + b_1 + \ldots + b_k)!}{a_1! \cdots a_k! b_1! \ldots b_k!} p_{I_1}^{a_1} \cdots p_{I_k}^{a_k} q_{I_1}^{b_1} \cdots q_{I_k}^{b_k}, \tag{12}$$

where $a_1 + \ldots + a_k + b_1 + \ldots + b_k = 2N$. Consequently, the joint distribution of the $2k$ variables $(Y_{(I_1,A)}, \ldots, Y_{(I_k,A)}, Y_{(I_1,B)}, \ldots, Y_{(I_k,B)})$ is a multinomial distribution.

The likelihood function of the distribution (Equation 12) is:

$$L(p_{I_1}, \ldots, p_{I_k}, q_{I_1}, \ldots, q_{I_k}) = \frac{(2N)!}{n_{I_1}! \ldots n_{I_k}! m_{I_1}! \ldots m_{I_k}!} p_{I_1}^{n_{I_1}} \ldots p_{I_k}^{n_{I_k}} q_{I_1}^{m_{I_1}} \ldots q_{I_k}^{m_{I_k}} \tag{13}$$

and since $p_{I_1} + \ldots + p_{I_k} + q_{I_1} + \ldots + q_{I_k} = 1$ it follows that the logarithm of this likelihood function remains as

$$\ln(L(p_{I_1}, \ldots, p_{I_k}, q_{I_1}, \ldots, q_{I_k})) = \ln\left(\frac{(2N)!}{n_{I_1}! \ldots n_{I_k}! m_{I_1}! \ldots m_{I_k}!}\right) + n_{I_1} \ln(p_{I_1}) + \ldots + n_{I_k} \ln(p_{I_k}) + m_{I_1} \ln(q_{I_1}) + \ldots + m_{I_m} \ln(1 - p_{I_1} - \ldots - p_{I_k} - q_{I_1} - \ldots - q_{I_{k-1}}).$$

In order to obtain the maximum likelihood estimators $\hat{p}_\alpha$ and $\hat{q}_\alpha$ of p_α, and q_α, respectively, for all $\alpha \in S$, we solve the following equations

$$\frac{\partial \ln(L(p_{I_1},\ldots,p_{I_k},q_{I_1},\ldots,q_{I_k}))}{\partial p_\alpha} = 0$$

$$\frac{\partial \ln(L(p_{I_1},\ldots,p_{I_k},q_{I_1},\ldots,q_{I_k})}{\partial q_\alpha} = 0 \tag{14}$$

for all $\alpha \in S$ to get that

$$\hat{p}_\alpha = \frac{n_\alpha}{2N}, \quad \hat{q}_\alpha = \frac{m_\alpha}{2N}, \tag{15}$$

and thus,

$$\hat{s}_\alpha = \frac{n_\alpha + m_\alpha}{2N}. \tag{16}$$

Then, under the null H_0, we have that $q_\alpha = p_\alpha$ and hence, $s_\alpha = p_\alpha + q_\alpha = p_\alpha + p_\alpha = 2p_\alpha$.

Therefore the likelihood ratio statistic is (see, e.g., Lehmann, 1986):

$$\lambda_i(Y) = \frac{p_{I_1}^{n_{I_1}} \cdots p_{I_k}^{n_{I_k}} q_{I_1}^{m_{I_1}} \cdots q_{I_k}^{m_{I_k}}}{\widehat{p}_{I_1}^{n_{I_1}} \cdots \widehat{p}_{I_k}^{n_{I_k}} \widehat{q}_{I_1}^{m_{I_1}} \cdots \widehat{q}_{I_k}^{m_{I_k}}} \tag{17}$$

$$= \frac{\left(\frac{1}{2}\right)^{2N} S_{I_1}^{n_{I_1}+m_{I_1}} \cdots S_{I_k}^{n_{I_k}+m_{I_k}}}{\left(\frac{n_{I_1}}{2N}\right)^{n_{I_1}} \cdots \left(\frac{n_{I_k}}{2N}\right)^{n_{I_k}} \left(\frac{m_{I_1}}{2N}\right)^{m_{I_1}} \cdots \left(\frac{m_{I_k}}{2N}\right)^{m_{I_k}}}. \tag{18}$$

On the other hand, $D(k) = -2\ln(\lambda_i(Y))$ asymptotically follows a chi-squared distribution with $k - 1$ degrees of freedom (see, for instance, Lehmann, 1986). Hence, we obtain that the estimator $\widehat{D}(k)$ of $D(k)$ is:

$$\begin{aligned}
\widehat{D}(k) &= -2\left[2N\ln\left(\frac{1}{2}\right) + (n_{I_1}+m_{I_1})\ln\left(\frac{n_{I_1}+m_{I_1}}{2N}\right) + \ldots\right. \\
&\quad + \ldots + (n_{I_k}+m_{I_k})\ln\left(\frac{n_{I_k}+m_{I_k}}{2N}\right) - n_{I_1}\ln\left(\frac{n_{I_1}}{2N}\right) - \ldots \\
&\quad \left. \ldots - n_{I_k}\ln\left(\frac{n_{I_k}}{2N}\right) - m_{I_1}\ln\left(\frac{m_{I_1}}{2N}\right) - \ldots - m_{I_k}\ln\left(\frac{m_{I_k}}{2N}\right)\right] \\
&= -4N\left[\ln\left(\frac{1}{2}\right) + \frac{n_{I_1}+m_{I_1}}{2N}\ln\left(\frac{n_{I_1}+m_{I_1}}{2N}\right) + \ldots\right. \\
&\quad \ldots + \frac{n_{I_k}+m_{I_k}}{2N}\ln\left(\frac{n_{I_k}+m_{I_k}}{2N}\right) - \frac{n_{I_1}}{2N}\ln\left(\frac{n_{I_1}}{2N}\right) - \\
&\quad \left. \ldots - \frac{n_{I_k}}{2N}\ln\left(\frac{n_{I_k}}{2N}\right) - \frac{m_{I_1}}{2N}\ln\left(\frac{m_{I_1}}{2N}\right) - \ldots - \frac{m_{I_k}}{2N}\ln\left(\frac{m_{I_k}}{2N}\right)\right] \\
&= 4N\left[\widehat{h}(S) - \widehat{h}(S,A) - \widehat{h}(S,B) - \ln\left(\frac{1}{2}\right)\right].
\end{aligned} \tag{19}$$

Therefore we have proved the following theorem.

Theorem 1. Let P be a population of cardinality N such that each individual $e \in P$ evaluates a fixed item. Assume that each individual gives his/her valuation in scales A and B. Denote by $h(S)$, $h(S, A)$, and $h(S, B)$ the total symbolic entropy of the valuation, of the valuation with scale A, and of valuation with scale B, respectively, as defined in Equation 3. If the valuation with scale A does not differ with the valuation with scale B, then

$$D(k) = 4N\left[h(S) - h(S,A) - h(S,B) - \ln\left(\frac{1}{2}\right)\right] \tag{20}$$

is asymptotically χ^2_{k-1} distributed.

Let α be a real number with $0 \le \alpha \le 1$. Let χ^2_α be such that

$$P\left(\chi^2_{k-1} > \chi^2_\alpha\right) = \alpha.$$

Then the decision rule in the application of the D-test at a $100(1-\alpha)\%$ confidence level is:

$$\begin{array}{l} \text{If } 0 \le D(k) \le \chi^2_\alpha \text{ Accept } H_0 \\ \quad \textit{Otherwise} \text{ Reject } H_0 \end{array}. \tag{21}$$

Properties of the *D*-test

Next we prove that the $D(k)$ test is consistent for a wide variety of alternatives to the null. This is a valuable property since the test will asymptotically reject that the valuation with scale A does not differ from the valuation with scale B whenever this assumption is not true. We will denote by $\widehat{D}(k)$ the estimator of $D(k)$. The proof of the following theorem can be found in Appendix.

Theorem 2. If the valuation with scale A differs from the evaluation with scale B, then

$$\lim_{N\to\infty} \Pr\left(\widehat{D}(k) > C\right) = 1 \tag{22}$$

for all $0 < C < \infty$, $C \in R$.

Since Theorem 1 implies $D(k) \to +\infty$ with probability approaching 1 whenever the valuations differ from each other, then upper-tailed critical values are appropriated.

It is important to note, from a practical point of view, that the researcher has to decide upon the parameter k in order to compute symbolic entropy and therefore, to calculate the $D(k)$ statistic. Fortunately, this decision can be easily conducted. Note that N should be larger than the number of symbols ($2k$).

Moreover, when the χ^2 is applied in practice, and all the expected frequencies are 5, the limiting tabulated χ^2 distribution gives, as a rule, the value χ^2_α with an approximation sufficient for ordinary purposes (see chap. 10 of Rohatgi (1976)). For this reason, we require to work with data sets containing *at least* $10k$.

Furthermore, we suggest to take k as the number of response alternatives of the scale with the lowest rank, namely l. The rationale of this decision is straightforward; by using this rule we ensure that all the symbols may occur. Otherwise if $k \ge l$, there might exist a symbol $\alpha \in S$ such that any valuation of the item in this scale does not belong to α and therefore α does not occur, while in the scale with more possible responses α occurs. This would lead to a rejection of the null even when it is true, in finite sample sizes.

Table 1. Results of the empirical application

Comparison[a]	*D*-test	Subintervals (k)	Chi-square cutoff values (95%)	Decision on H_0
A–B	25.26	5	9.48	Reject
A–C	12.49	7	12.59	Accept
A–D	2.42	7	12.59	Accept

Note. [a]Free-scale (A); Likert-type scale from 1 to 5 (B); Likert-type scale from 1 to 7 (C); Semantic-differential scale from −3 to +3 (D).

Empirical Application

We applied *D*-test to a real marketing research. The aim of the study was to test if scale invariance holds when measuring service quality in a sports service. Measuring service quality has been a very relevant topic in the academic marketing literature during the last three decades. The advent of international quality standards, such as the International Organisation for Standards (ISO) 9000 series and the European Foundation for Quality Management (EFQM) model, has been a significant development in quality management.

We collected a random sample of 212 customers of a sports center placed in a Spanish city. Data were collected during the spring of 2008 by personal interview. Business and marketing students had been previously trained for this task. Individuals had to evaluate service quality using a linguistic term and then a numeric value. When they indicated the numeric value, they had to indicate in which scale this value had meaning. This was taken as the free-scale. Consequently, responses would not be restricted to a categorical scale proposed by the researcher. In this situation, individuals responded in the scale that fitted with their preferences, minimizing their psychological costs (Ferrando, 2003; Weng & Cheng, 2000).

Then, after two questions regarding other variables, the interviewer returned to inquire about service quality using the following expression: "*You have indicated that your service quality perception was XXX* (the interviewer used the same linguistic term that the consumer had used previously), *so how would you represent that valuation on the following scales?*" (semantic-differential scale from −3 to +3, Likert-type scale from 1 to 5 and from 1 to 7). Therefore, by separating the questions about service quality in groups, we tried to minimize the possibility of appearance of dependency between numerical responses. We wanted to get a conditional independent model, as Figure 1 illustrates.

Consumers preferred to give their responses in continuous scales: from 1 to 10, and from 0 to 10. The percentage of responses was 63.2 and 36.8, respectively. We say "continuous" because they often used decimal numbers to evaluate service quality. All these responses were transformed to a unit [0,1] interval. This was the free-scale. The remaining data derived from the other rating scales were also transformed.

We applied *D*-test running a simple *Mathematica 6.0* program. This program is available upon request. We achieved several pairwise comparisons between the responses derived from the free-scale and the other three rating scales. Results are shown in Table 1. Then, we reject the hypothesis of scale invariance for the comparison between the free-scale and the Likert-type scale from 1 to 5. However, we cannot reject the hypothesis in the remaining comparisons. Consequently, scale invariance holds when responses are taken in a 7-point scale (Likert-type scale from 1 to 7 and semantic-differential scale from −3 to +3).

Implications for Marketing Research

We have constructed a new, simple, nonparametric, consistent, and powerful test to analyze scale invariance in marketing research. The use of rating scales in marketing is very widespread, consequently *D*-test provides researchers a tool to decide the appropriate rating scale to study specific marketing problems, and how the results of prior studies can be questioned. Several implications deserve to be highlighted:

Firstly, scale invariance has to be a necessary condition to use a categorical rating scale, in order to make the categorization error negligible. We provide a test that overcomes the shortcomings of other procedures for analyzing this important concern. *D*-test requires fewer assumptions than the direct comparison method or structural equation modeling. The main requirement of *D*-test is to work with data sets containing at least $10k$, that is, sample size has to be 10 times the number of subintervals. As the number of subintervals has to be fixed by the researcher to the rank of the scale with the fewest number of alternatives of response, then this requirement is not very demanding.

Secondly, we stress the necessity of replication. Scale invariance should be studied in other services and using other attitudinal variables. This is a necessary condition to provide a general recommendation about the effect of measuring attitudinal variables. By knowing these effects, meta-analysis should be reconsidered. For example, are two Cohen's *d* effect sizes commensurable when they come from two distinct rating scales?

Thirdly, verbal labels may interact with numerical responses (Javaras & Ripley, 2007; Saris & Gallhofer, 2007). Note that we have compared the free-scale with three rating scales without verbal labels. It would be very interesting for further research to accomplish the same study using verbal labels for the rating scales. As interaction may occur, *D*-test could yield different results, and probably the hypothesis of scale invariance would be rejected with more assiduity. This research scenario would be more realistic, because almost all rating scales used in management and marketing

research are labeled with verbal terms (good-bad, very satisfied-very unsatisfied, strongly agree-strongly disagree, etc.). Moreover, further research should deepen into the interaction of verbal labels with numerical responses using the fuzzy logic approach.

Fourthly, it is important to keep in mind that the data are originated from the interaction of the scale and the person responding on the scale. This means that in another sample, the conclusions can be totally different. Especially this feature is why people should test for scale invariance in each sample separately, and why the *D*-test is of crucial importance.[3]

Finally, findings from our empirical example have showed that consumers in this specific sample used a common framework to evaluate service quality (the 10-point scales). It seems plausible that this type of scale will also be preferred for judging other related variables, as, for example, consumer satisfaction. We might think data that come from these scales as continuous, and this could lead researchers to reconsider some assumptions of the latent variable approach, that is attitudes are latent variables but probably they are not "scale-free." We mean that in some cultural contexts, attitudes may be continuously distributed in a restricted rank in the mind of the consumer: from 0 or 1 to 10, and they are not varying in the $(-\infty, \infty)$ interval. Fortunately, *D*-test can successfully deal with a mixture of continuous and categorical data because of its nonparametric nature.

Acknowledgments

Manuel Ruiz was partially supported by MEC (Ministerio de Educación y Ciencia), Grant MTM2009-07373 and Fundación Séneca of Región de Murcia. Jose A. Martínez was partially supported by Fundación Séneca of Región de Murcia. Authors acknowledge editor and three anonymous reviewers for their helpful suggestions. In addition, we are indebted to Joeri Hofmans for his helpful comments in previous versions of this article.

References

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 11*, 1061–1071.

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

Ferrando, P. J. (2003). A Kernel density analysis of continuous typical-response scales. *Educational and Psychological Measurement, 63*, 809–824.

Hayduk, L. A. (1996). *LISREL: Issues, debates and strategies*. Baltimore, MA: Johns Hopkins University Press.

Hofmans, J., Theuns, P., Baekelandt, S., Mairesse, O., Schillewaert, N., & Cools, W. (2007). Bias and changes in perceived intensity of verbal qualifiers effected by scale orientation. *Survey Research Methods, 1*, 97–108.

Hofmans, J., Theuns, P., & Mairesse, O. (2007). On the impact of the number of response categories on linearity and sensitivity of 'Self Anchoring Scales'. A Functional Measurement approach". *Methodology, 3*, 160–169.

Jackson, D. L. (2003). Revisiting sample size and number of parameter estimates: Some support for the N:q hypothesis. *Structural Equation Modeling, 10*, 128–141.

Javaras, K. N., & Ripley, B. D. (2007). An "unfolding" latent variable model for Likert attitude data: Drawing inferences adjusted for response style. *Journal of the American Statistical Association, 102*, 454–463.

Jöreskog, K. G., & Sörbom, D. (2001). *LISREL 8: user's reference guide*. Scientific Software International.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*, 263–291.

Lehmann, E. L. (1986). *Testing Statistical hypothesis*. New York, NY: Wiley.

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

Lindstrom, M. (2008). *Buyology*. New York, NY: Doubleday.

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

Mast, F. W., & Zaltman, G. (2005). A behavioral window on the mind of the market: An application of the response time paradigm. *Brain Research Bulletin, 67*, 422–427.

Matilla, M., & Ruiz, M. (2008). A non-parametric independence test using permutation entropy. *Journal of Econometrics, 144*, 139–155.

Rohatgi, V. K. (1976). *An introduction to probability theory and mathematical statistics*. New York, NY: Wiley.

Saris, W. E., & Gallhofer, I. (2007). Estimation of the effects of measurement characteristics on the quality of survey questions. *Survey Research Methods, 1*, 29–43.

Tomarken, A. J., & Waller, N. G. (2005). Structural equation modeling: Strengths, limitations, and misconceptions. *Annual Review of Clinical Psychology, 1*, 31–65.

Weng, L.-J., & Cheng, C.-P. (2000). Effects of response order on Likert-type scales. *Educational and Psychological Measurement, 60*, 908–924.

Yuan, K. H., & Bentler, P. M. (2004). On chi-square difference and *z*-tests in mean and covariance structure analysis when the base model is misspecified. *Educational and Psychological Measurement, 64*, 737–757.

Jose A. Martínez and Manuel Ruiz Marín

Facultad de Ciencias de la Empresa
Universidad Politécnica de Cartagena
Paseo Alfonso XIII, 50
30203 Cartagena
Spain
Tel. +34 968 32 57 76
Fax +34 968 32 70 81
E-mail josean.martinez@upct.es, manuel.ruiz@upct.es

[3] We thank a Reviewer for this comment.

Appendix

Proofs

Proof of Theorem 2

Proof. First notice that the estimators

$$\widehat{h}(S,A) = -\sum_{\alpha\in S}\widehat{p}_\alpha \ln(\widehat{p}_\alpha)$$
$$\widehat{h}(S,B) = -\sum_{\alpha\in S}\widehat{q}_\alpha \ln(\widehat{q}_\alpha)$$
$$\widehat{h}(S) = -\sum_{\alpha\in S}\widehat{s}_\alpha \ln(\widehat{s}_\alpha) \tag{23}$$

of $h(S, A)$, $h(S, B)$ and $h(S)$, respectively, where $\widehat{p}_\alpha = \frac{n_\alpha}{2N}$, $\widehat{q}_\alpha = \frac{m_\alpha}{2N}$, and $\widehat{s}_\alpha = \frac{n_\alpha+m_\alpha}{2N}$ are consistent because $p\lim_{N\to\infty} \widehat{p}_\alpha = p_\alpha$, $p\lim_{N\to\infty} \widehat{q}_\alpha = q_\alpha$, and $p\lim_{N\to\infty} \widehat{s}_\alpha = s_\alpha$ and hence

$$p\lim_{N\to\infty} \widehat{h}(S,A) = h(S,A)$$
$$p\lim_{N\to\infty} \widehat{h}(S,B) = h(S,B)$$
$$p\lim_{N\to\infty} \widehat{h}(S) = h(S) \tag{24}$$

Denote by $\Delta(k) = h(S) - h(S,A) - h(S,B) - \ln(\frac{1}{2})$ and recall that

$$\Delta(k) = -\sum_{\alpha\in S}(p_\alpha + q_\alpha)\ln(p_\alpha + q_\alpha) + \sum_{\alpha\in S}(p_\alpha)\ln(p_\alpha) + \sum_{\alpha\in S}(q_\alpha)\ln(q_\alpha) - \ln\left(\frac{1}{2}\right) \tag{25}$$

Since $\sum_{\alpha\in S} p_\alpha = \frac{1}{2}$, under the null, that is, when $p_\alpha = q_\alpha$ for all $\alpha \in S$, it follows that

$$\begin{aligned}\Delta(k) &= -2\sum_{\alpha\in S} p_\alpha \ln(2p_\alpha) + 2\sum_{\alpha\in S} p_\alpha \ln(p_\alpha) - \ln\left(\frac{1}{2}\right)\\ &= -2\sum_{\alpha\in S} p_\alpha \ln(2) - 2\sum_{\alpha\in S} p_\alpha \ln(p_\alpha) + 2\sum_{\alpha\in S} p_\alpha \ln(p_\alpha) - \ln\left(\frac{1}{2}\right)\\ &= -\ln(2) - \ln\left(\frac{1}{2}\right) = 0\end{aligned} \tag{26}$$

Otherwise, that is, when H_1 holds, and taking into account that $-\ln(x) > 1 - x$ we have that

$$\begin{aligned}\Delta(k) &= -\sum_{\alpha\in S}(p_\alpha + q_\alpha)\ln(p_\alpha + q_\alpha) + \sum_{\alpha\in S}(p_\alpha)\ln(p_\alpha) + \sum_{\alpha\in S}(q_\alpha)\ln(q_\alpha) - \ln\left(\frac{1}{2}\right) >\\ &= \sum_{\alpha\in S}(p_\alpha + q_\alpha)(1 - p_\alpha - q_\alpha) - \sum_{\alpha\in S} p_\alpha(1-p_\alpha) - \sum_{\alpha\in S} q_\alpha(1-q_\alpha) + \frac{1}{2}\\ &= -\sum_{\alpha\in S}(p_\alpha + q_\alpha)^2 + \sum_{\alpha\in S} p_\alpha^2 + \sum_{\alpha\in S} q_\alpha^2 + \frac{1}{2}\\ &= -2\sum_{\alpha\in S} p_\alpha q_\alpha + \frac{1}{2}.\end{aligned} \tag{27}$$

Now, since $\sum_{\alpha\in S} p_\alpha = \frac{1}{2} = \sum_{\alpha\in S} q_\alpha$ we have that

$$\frac{1}{4} = \sum_{\alpha\in S} p_\alpha \sum_{\alpha\in S} q_\alpha = \sum_{\alpha\in S} p_\alpha q_\alpha + \sum_{\alpha\neq\beta} p_\alpha q_\beta \geq \sum_{\alpha\in S} p_\alpha q_\alpha \tag{28}$$

and thus,

$$2\frac{1}{4} = \frac{1}{2} \geq 2\sum_{\alpha\in S} p_\alpha q_\alpha. \tag{29}$$

Therefore, under H_1, by (27) and (29) we get that $\Delta(k) > 0$.

Let $0 < C < \infty$ with $C \in \mathbb{R}$ and take N large enough such that

$$\frac{C}{2N} < D(k). \tag{30}$$

Then since $D(k) = 2N\Delta(k)$ it follows that

$$\Pr[\widehat{D}(k) > C] = \Pr\left[\widehat{\Delta}(k) > \frac{C}{2N}\right] \tag{31}$$

Therefore, by (30) and (31) we get that

$$\lim_{N\to\infty} \Pr\left(\widehat{D}(k) > C\right) = 1$$

as desired. □