A Method to Analyse Measurement Invariance under Uncertainty in Between-Subjects Design

Article *in* The Spanish Journal of Psychology · November 2012

DOI: 10.5209/rev_SJOP.2012.v15.n3.39435 · Source: PubMed

Jose A. Martinez
Universidad Politécnica de Cartagena
139 PUBLICATIONS **2,009** CITATIONS

Manuel Ruiz Marín
Universidad Politécnica de Cartagena
136 PUBLICATIONS **1,117** CITATIONS

The Spanish Journal of Psychology
2012, Vol. 15, No. 3, On Line First
http://dx.doi.org/10.5209/rev_SJOP.2012.v15.n3.39435

ISSN 1138-7416

A Method to Analyse Measurement Invariance under Uncertainty in between-Subjects Design

José A. Martínez[1], Manuel Ruiz Marin[1], and Maria del Carmen Vivo Molina[2]

[1]Universidad Politécnica de Cartagena (Spain)
[2]Fundación para la Formación e Investigación Sanitaria de la Región de Murcia (Spain)

In this research we have introduced a new test (H-test) for analyzing scale invariance in between group designs, and considering uncertainty in individual responses, in order to study the adequacy of disparate rating and visual scales for measuring abstract concepts. The H-test is easy to compute and, as a nonparametric test, does not require any a priori distribution of the data nor conditions on the variances of the distributions to be tested. We apply this test to measure perceived service quality of consumers of a sports services. Results show that, without considering uncertainty, the 1-7 scale is invariant, in line with the related works regarding this topic. However, de 1-5 scale and the 1-7 scale are invariant when adding uncertainty to the analysis. Therefore, adding uncertainty importantly change the conclusions regarding invariance analysis. Both types of visual scales are not invariant in the uncertainty scenario. Implications for the use of rating scales are discussed.

Keywords: measurement invariance, between groups design, rating scales, uncertainty, H-test.

En esta investigación presentamos un nuevo test (test H) para analizar la invarianza de escala en diseños entre sujetos, considerando además la incertidumbre en las respuestas de los individuos, con el fin de estudiar la idoneidad de diferentes escalas de medición de conceptos abstractos. El test H es fácil de calcular y, debido a su naturaleza no paramétrica, no requiere ninguna asunción a prior sobre la distribución de los datos ni de las condiciones de la varianza. Aplicamos este test para medir la calidad percibida de los consumidores de servicios deportivos, y los resultados muestran que, sin considerar la incertidumbre, la escala de 1 a 7 es invariante, en línea con las conclusiones obtenidas en otras investigaciones. Sin embargo, al añadir la incertidumbre en el análisis, las escalas de 1 a 5 y de 1 a 7 son invariantes. Por tanto, la consideración de la incertidumbre cambia las conclusiones en relación al análisis de la invarianza de escala. Las escalas visuales consideradas, a su vez, no son invariantes. Finalmente, las implicaciones para el uso de escalas de medición son discutidas.

Palabras clave: invarianza de escala, diseño entre grupos, escalas de medición, incertidumbre, test H.

Manuel Ruiz Marín has been partially supported by the project MTM2009-07373 from Ministerio de Ciencia e Innovación of Spain, Fundación para la Formación e Investigación Sanitarias and Fundación Seneca of the Región de Murcia.

Correspondence concerning this article should be addressed to José Antonio Martínez García. Departamento de Economía de la Empresa, Facultad de Ciencias de la Empresa, Universidad Politécnica de Cartagena. Paseo Alfonso XIII, 30203 Cartagena – Murcia (Spain). Phone: +34-968325776. E-mail: josean.martinez@upct.es

Rating scales are widely used in psychology. A great amount of studies covering disparate topics such as motivation (Gregg & Hall, 2006); personality (Amigó, Caselles, & Micó, 2010), burnout (Creswell & Eklund, 2006), talent development (Martindale et al., 2010), life satisfaction (San Martín, Perles, & Cantó, 2010), consumer perceptions (Ko & Pastore, 2005) etc. use these type of measurement instruments. These scales are mainly used to measure subjective opinions/perceptions/attitudes of research participants, i.e. self-report measures of different variables.

The use of rating scales is mainly achieved under the "third person" approach, i.e. where a researcher proposes a specific length of the rating scale to the respondent. This contrasts with the use of the first-person approach for measurement (Kilpatrick & Cantril, 1960; Zaltman & Zaltman, 2008), i.e. participants would respond using the "best scale" (that is, the scale that fits with their preferences, also called the "free scale") within the context of the question and the answer, i.e. using the scale that minimizes their psychological costs. This possible gap between a respondent preferred scale and the scale he/she has to use to complete the questionnaire may create bias due to categorization error (Cox, 1980). This type of error occurs when a respondent has to convert the value he/she assign to the variable of interest into a category that does not match with that value, i.e. when the length and/or the numeric labels of both scales (the free scale and the proposed scale) differ. However, this bias is negligible if the proposed scale is invariant. Scale invariance is a unifying psychological principle, and under this setting we will say that the distribution of a magnitude is observed to be scale invariant if the statistical structure remains the same at different measurement scales, i.e. if the measure of the variable of interest does not change if the length of the scale is multiplied by a constant factor; for example, when the free-scale is a 10-point scale and the proposed scale in the questionnaire is a 5-point scale.

As the first-person approach is underused in applied research, several recent studies have treated to analyse if some specific rating scales are invariant. Therefore, the studies of Martínez, Ko, and Martínez (2010), and Martínez and Ruiz (2011) analyse the use of rating scales for evaluating quality and satisfaction of consumers with sports services, and propose the D-test, a test for studying scale invariance using symbolic dynamics and symbolic entropy, respectively. These studies analyse three disparate rating scales (from 1 to 5, from 1 to 7, and from -3 to +3) to compare with the free-scale, in order to know if scale invariance holds. Taking together, these two studies show that the two seven point scales are invariant but not the five point scale. Therefore, if the third person approach for measurement is used in this research context, these two seven point scales can be used without bias, but not the five point scale. In addition, Martínez et al. (2010) studied the relationship between this optimized numerical response and a verbal label associated with this number, considering the uncertainty of the verbal responses, how individuals use these verbal labels, and how these responses are distributed, using the fuzzy logic approach.

However, there are still several problems which should be addresses, in order to advance in this stream of research. First of all, the D-test (Martínez & Ruiz, 2011) proposed to analyse scale invariance, can only be applied when the same participant valuate the variable of interest using disparate rating scales. Therefore, one of the shortcomings of the D-test is the possible dependence of responses. Recall that when a participant evaluated service quality of a sport facility in a free-scale, then had to evaluate the same variable using other rating scales, so principle of conditional independence (Hayduk, 1996) could be violated. Second, although the three rating scales analysed (from 1 to 5, from 1 to 7, and from -3 to +3) are probably the most used scales in social sciences, it would be interesting to expand the invariance analysis to other type of scales. And third, although Martínez et al. (2010) consider the uncertainty of responses of participants, they aggregate individual responses to form triangular fuzzy numbers. Therefore, they do not consider individual uncertainty, and ultimately, they do not take into account uncertainty in the analysis of scale invariance.

The aim of this research is to solve these three commented limitations of the existing research by the following way: (1) we propose a non-parametric test to analyse scale invariance which may be applied to independent groups. Our new test overcome this limitation, because participants only give their responses using a single scale, and responses of these different group of participants (each group using a different scale), may be compared. The proposed test does not assume any distribution of the data and no restriction on the variance of the variables is needed. Moreover, the new test seems to detect different distributions by testing on the median, where the usual ANOVA and t-test accept the hypothesis of equal means (not necessarily equal distributions); (2) we add two new scales to the invariance analysis: a graphic scale visualized in a computer, and a graphic scale presented in a paper. Participants had to use a digital and a normal pen, respectively, to give their responses. Graphic scales are also used in social sciences (Shamir & Kark, 2004), as an alternative to the traditional rating scales, and they have the advantage to deal with imprecision of responses in a more easy fashion. As such, most of them have the form of a vertical or horizontal line, on which the respondent is asked to place his or her response between two poles that represent the extreme points of the attitudes being measured. As Shamir and Kark (2004) remember, quoting the work of Gardner, Cummings, Dunham, and Pierce (1998), most standard questionnaires measure psychological constructs by using multiple-item Likert-type verbal scales that have a similar appearance and a like response options. For a diversity of reasons, including monotony and response sets, the relationships between these constructs can be partly attributed to common method variance; (3) we consider the uncertainty of responses of participants in our statistical test of scale invariance. Participants could indicate if their responses had imprecision using the three rating scales and the free-scale. In addition, graphic scales implicitly counted for imprecision. Adding imprecision to the individual responses may change observed responses, and then statistical analysis. We use the fuzzy logic reasoning, considering the triangular distribution as a form to model individual uncertainty

Therefore, the relevance of this research is two-fold: First, we propose a new, powerful and simple non-parametric test to assess measurement invariance which can be used to evaluate any rating or graphic scale; Second we apply the new test to provide robust results to the analysis of measurement invariance considering uncertainty in responses in an important field of consumer psychology, such as perceived service quality research.

Measurement error and uncertainty

Classical Test Theory postulates that the observable measures are related with true scores by the following equation (1):

$$x_i = \mathrm{a} + bX_i + e_i \qquad (1)$$

Where x is the observable measure, X is the true score, a,b are constants parameters and e is the measurement error. Given the arbitrariness of scales in psychology, and if we normalize all measures in a [0,1] interval (see Cohen, Cohen, Aiken & West, 1999), then the prior equation may be simplified to (2):

$$x_i = X_i + e_i \qquad (2)$$

where exists a simple linear and unitary relationship between the observable and the true score. This linear assumption is often considered as close enough to the more complex non-linear relationship assumed by Item Response Theory (see Schmidt & Hunter, 1996).

It seems clear that if $e_i \approx IIDF(0,\sigma^2)$, where F is an arbitrary distribution function, then $E(x_i) = X$ and $Var(x_i) = \sigma^2$.

Therefore, the key factors to obtain non-biased observable measures of an operationally defined latent variable, is to guarantee that there are non-systematic variables included in the error term, i.e., that variation in the observable measure is purely random.

This is the most common scenario in cross-sectional research, where Equation (4) is a classical regression equation (Spanos, 2010). Again, error term should be $e_i \approx IIDF(0,\sigma^2)$.

Schmidt and Hunter (1999) speak about different sources of measurement error: random, transient and specific. These types of error could be considered white noise under the absence of other threats to measurement. One of these threats is specifically acknowledged by Schmidt and Hunter (1996): categorization error. This type of error occurs when a quantifiable measure is codified to a discrete category, where there is no perfect match between the original measure and the assigned category. Scherpenzeel and Saris (1997) show how to specify this error in the context of a method effect. In addition, this is explained perfectly by Cox (1980), and it is a problem when individuals prefer to give their responses in a specific scale, and researchers only give to them the option to answer in a different scale. This is the problem of scale invariance which has been studied by Martínez and Ruiz (2011). Therefore, in order to obtain non-biased observable measures, scale invariance must hold.

There is, however, another issue about measurement that psychometrics has not traditionally treated: uncertainty. We do not refer to the change in observable responses due to transient or random error. We refer to the fact that the operationally defined latent variable would not be a fixed specific value, but a range of values due to the uncertainty of individual thoughts. For example, this occurs when individuals assign a range of numerical values to a verbal label representing their thoughts (Hersh & Caramazza, 1974), or when they define the same linguistic concept using disparate terms (Nicolai & Dawitz, 2009). Likewise, when, for example, an individual says that is very satisfied with the service provided by a sport centre, this may mean that assign a range of numeric values coded under the same verbal term. This issue has been widely treated in other field outside psychometrics, as operation research, through the fuzzy logic approach Martínez et al. (2010).

Including uncertainty transform the Equation (2) in the following expression (3):

$$x_i^* = X_i^* + e_i^* \quad (3)$$

Where all these variables are not single numbers (crisp numbers in the fuzzy logic terminology), but a range of possible numeric values that are characterized by a specific distribution. Researchers often use the triangular distribution Martínez et al. (2010) for that purpose.

In this research we propose a new and simple form to consider uncertainty under the Classical Test Theory of measurement. We convert triangular fuzzy numbers to crisp numbers in order to analyse categorization error in individual responses under different measurement scales. Therefore, we provide a framework to analyse scale invariance under uncertainty in sports sciences, with the aim to maximize utility of individual responses.

Method

We designed an experiment creating six different conditions corresponding with the six different measurement scales. Participants were randomly assigned to each one of the six treatments. The procedure to obtain the data was the following: We announced in our university that we were achieving a marketing experiment with undergraduate students. In this announcement, we demanded the voluntary participation of the alumni in order to answer a few simple questions using a computer. To promote participation we communicated we raffled an iPOD Nano (valued at 125 Euros) among all participants. We prepared a room with a printer, scanner, and a computer with a Wacom tablet in order to allow participants to respond to the questionnaire using a digital pen. A specialized computing company created the digital questionnaire following our indications. As there was only one computer, participants had to wait outside the room till the prior student finish the experiment. Although it was a self-administered questionnaire, another trained graduate student was always in the room in order to solve any doubt. Finally, 202 students completed the questionnaire.

Measures

Similar to the study of Martínez et al. (2010), participants had to respond to the following simple question: "*Please, indicate your perception about the service quality provided by public sports services in your city*". For the control group, participants had to respond using the scale they preferred. They had to respond using a number and then to indicate the reference system for that number. For the remaining five groups, participants had to respond using the following scales: from -3 to +3, from 1 to 5, from 1 to 7, a visual scale showed in the computer, and a visual scale showed in a paper, respectively. For both visual graphic scales, we showed three lines with different length (50, 75 and 100 mm). Participants had to freely choose one of the three alternatives and then give their response using a digital pen for the computer treatment, or a normal pen to the paper treatment. Responses in the paper were subsequently scanned and incorporated to the computed data base. No verbal labels were indicated in any of the scales, in order to avoid interaction between verbal and numerical responses (Martínez et al., 2010).

Once participants had indicated their service quality perception, then they had to indicate if their responses had uncertainty. We showed an example in the computer about what uncertainty meant. If they answered "Yes", then they had to indicate the range of uncertainty in the specific scale they were using. For the visual scales, uncertainty was implicitly derived from the width of the responses. For the visual scale showed in the computed, pixels length were 156, 234 and 312 for the lines of 50, 75 and 100 mm, respectively. Sensibility of the digital pen was about 1 pixel. Considering the width of the 17" screen (1204x768 pixels), then a tolerance of .32 mm was permitted. Therefore, when participants responded using a width higher than .32 mm, then uncertainty was present. Regarding the visual scale showed in the paper, pixels length were higher because scanner had more resolution than the screen. Then pixels length were 246, 443 and 590, for the lines 50, 75 and 100 mm, respectively. Scanner tolerance was on 5 pixels (.8 mm), and pen sensibility was on .8 mm, therefore when participants responded using a width higher than 1.6 mm, then uncertainty was present.

The following question was on their preferred scale. Participants assigned to the five experimental conditions had to indicate if they preferred to give their responses using another measurement scale. Finally, participants had to respond to a few socio-demographical questions.

Measuring the uncertainty

Assume that each individual in the population is asked, after the valuation of the item I, about the uncertainty of his valuation. That is to say, each individual, namely e gives his valuation x_e of the item I and an interval (a_e,b_e) in which his valuation would fluctuate due to uncertainty.

Under these assumptions one may assume that the distribution of the valuation of the item I by individual e takes the value 0 out of the interval (a_e,b_e) and gets its maximum in x_e.

Therefore we may assume that this distribution is a triangular distribution with lower limit a_e, mode x_e and upper limit b_e. More specifically, the probability density function of this triangular distribution is given by

$$f(k)=\begin{cases}\dfrac{2(k-a_e)}{(b_e-a_e)(x_e-a_e)} & \text{if } a_e\le k\le x_e\\ \dfrac{2(b_e-k)}{(b_e-a_e)(b_e-x_e)} & \text{if } x_e\le k\le b_e\end{cases}$$

It is straightforward to check that the expected value of the triangular distribution is

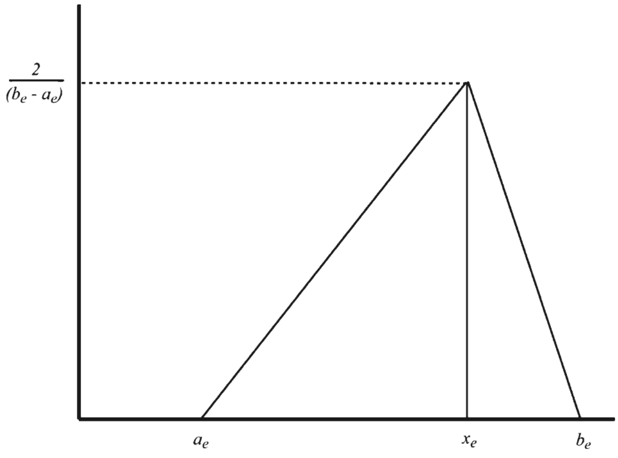

Figure 1. Example of a triangular fuzzy number.

$$\mu_e = \frac{a_e + x_e + b_e}{3}$$

its mode is $mo_e = x_e$, its median is

$$Me = \begin{cases} x_e & \text{if } x_e = \dfrac{b_e + a_e}{2} \\ a_e + \sqrt{\dfrac{(b_e - a_e)(x_e - a_e)}{2}} & \text{if } x_e \geq \dfrac{b_e + a_e}{2} \\ b_e - \sqrt{\dfrac{(b_e - a_e)(b_e - x_e)}{2}} & \text{if } x_e \leq \dfrac{b_e + a_e}{2} \end{cases}$$

and its variance is

$$\sigma_e = \frac{a_e^2 + x_e^2 + b_e^2 - a_e b_e - a_e x_e - b_e x_e}{18}\ .$$

Results

First of all we depict the responses of the participants in the free-scale group. Table 1 shows that 0-10 and 1-10 scale are the scales preferred by participants. The other scale types are marginally represented, but they are a sign of the heterogeneity of individual reference systems. It is also highly noticeable that none individual indicated the other rating scales considered in this study as their preferred scale.

In the following step, we depict the descriptive statistics of each one of the groups under both scenarios of analysis: non-uncertainty and uncertainty (Table 2). In order to be able to compare the valuations among different scales we have normalized the responses to the unit interval (see next section for a more detailed explanation on the normalization procedure). In the case of uncertainty the valuation of the item is given by the expected value of the associated triangular distribution of each individual. Recall that in both visual groups the scores are the same, because when we computed crisp scores from the fuzzy graphical responses, we applied the triangular transformation depicted before.

Valid responses diminished in the uncertainty case, because some individuals failed to give a proper response once admitted they were uncertain in their service quality perceptions. A total of 12 individuals correctly answered the uncertainty questions for the non-visual scales. Therefore, only a slight part of the sample admitted uncertainty in their judgements. However, as we show afterwards, these small changes influence invariance analysis.

Before using the t-test in order to detect mean differences between each of the scales and the free scale we have tested for normality. In order to do so we used the Kolmogorov-Smirnov test.

As can be seen in Table 3 all scales distribute as a Normal distribution except for the 1-5 scale in the case of no uncertainty where we will accept normality with a p-value of .952.

Table 4 shows the mean difference confidence interval together with the two sided p-value of the mean difference test for the free scale and each one of the remaining 5 scales with

Table 1
Scale type preferred for the participants in the free-scale group

	Scale type								
Valid responses	0-10	0-100	1-10	3-6	4-6	4-7	5-8	60-75	5-15
30	53.33%	6.66%	16.66%	3.33%	6.66%	3.33%	3.33%	3.33%	3.33%

Table 2
Descriptive statistics

	Scale type					
No uncertainty	Free	1-5	–3 +3	1-7	Visual (computer)	Visual (paper)
Valid responses	30	35	34	31	35	32
Mean	.547	.507	.608	.473	.496	.513
Std. deviation	.152	.230	.204	.243	.187	.232
Uncertainty						
Valid responses (responses with uncertainty)	29 (6)	29 (4)	31 (1)	28 (1)	35 (35)	31 (31)
Mean	.561	.508	.625	.462	.496	.522
Std. deviation	.139	.240	.205	.248	.187	.231

Table 3
Kolmogorov-Smirnov test

	Scale type					
No uncertainty	Free	1-5	–3 +3	1-7	Visual (computer)	Visual (paper)
Statistic	1.105	1.364	1.036	0.964	1.063	0.738
Bilateral signification	.174	.048	.234	.310	.208	.648
Uncertainty						
Statistic	0.869	1.130	1.074	0.862	0.812	1.033
Bilateral signification	.437	.156	.199	.447	.525	.237

Table 4
Mean difference 95% confidence interval and two sided p-value

Scale type	No Uncertainty	1-5	–3 +3	1-7	Visual (Computer)	Visual (Paper)
Free	CI 95%	(–.136,.055)	(–.028,.150)	(–.178,.030)	(–.135,.033)	(–.133,.065)
	Two sided p-value	.405	.178	.159	.231	.498
Free Uncertainty	CI 95%	(–.147,.039)	(–.040,.134)	(–.190,.014)	(–.146,.017)	(–.145,.049)
	Two sided p-value	.254	.284	.090	.331	.118
	No Uncertainty	1-5	–3 +3	1-7	Visual (Computer)	Visual (Paper)
Free	CI 95%	(–.144,.067)	(–.014,.170)	(–.195,.024)	(–.134,.033)	(–.125,.075)
	Two sided p-value	.465	.096	.126	.617	.234
Free Uncertainty	CI 95%	(–.156,.051)	(–.026,.154)	(–.207,.009)	(–.146,.017)	(–.137,.059)
	Two sided p-value	.314	.158	.072	.430	.121

and without uncertainty. In all cases the test does not reject the null hypothesis of equal means at a 95% confidence level.

Notice that equal means does not implies equal distributions. In order to check if the distribution of the free scale coincides with the distribution of the other scales we have developed the following nonparametric test.

Construction of the test

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, namely item I. In order to do so every individual has to chose between two different scales A and B. Let P_A and P_B be the two subsets of P such that every element in P_A chooses scale A to evaluate the item I and every element in P_B chooses scale B to evaluate the item I. Denote by n_A the cardinality of population P_A and n_B the cardinality of population P_B.

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:

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

Denote by Me_A and Me_B the median of the distribution of the valuation of item I in scales A and B respectively once they are normalized to the unit interval. For any individual $e\in P_A$ let

$$I_A(e) = \begin{cases} 1 & \text{if } \tilde{x}_e \le Me_B \\ 0 & \text{otherwise.} \end{cases} \qquad (1)$$

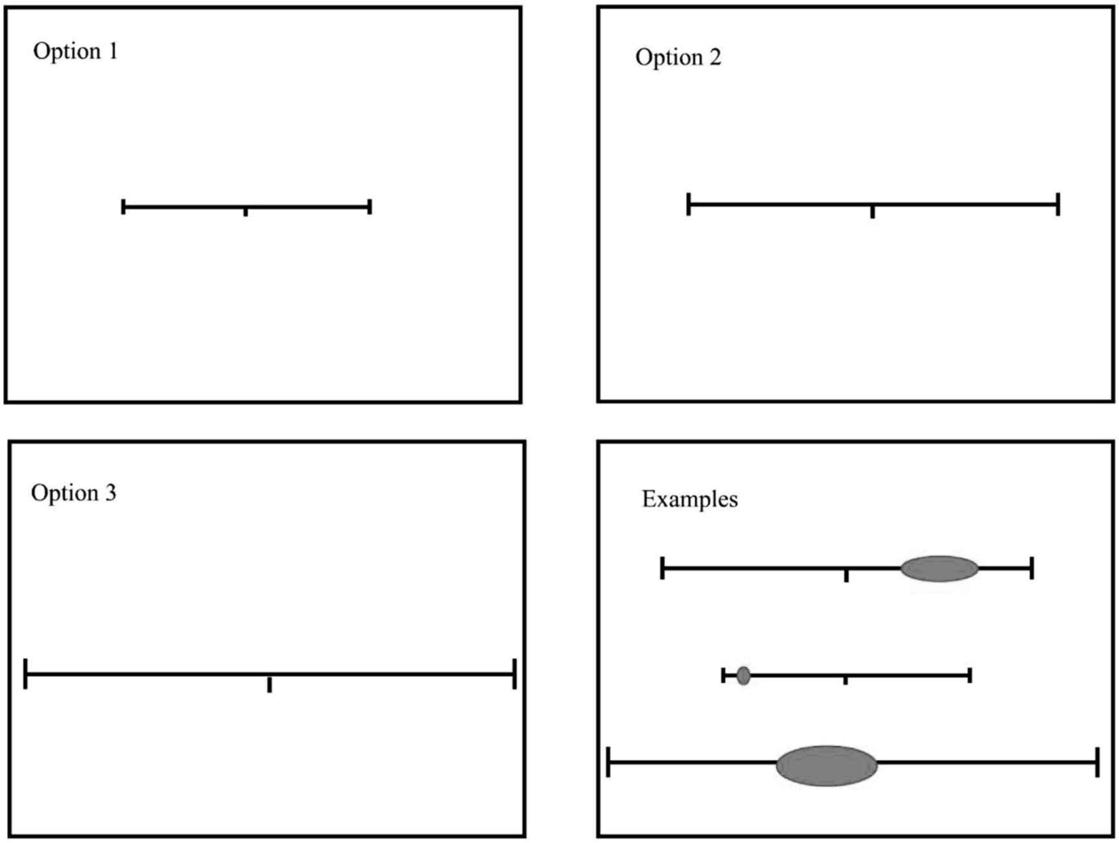

Figure 2. Display screen to sample 4

Similarly for an individual $e \in P_B$ let

$$I_B(e) = \begin{cases} 1 & \text{if } \tilde{y}_e \leq Me_A \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

We will denote by p_A (respectively p_B) the probability for an individual in population P_A (respectively P_B) to have a valuation of item I smaller than Me_B (respectively Me_A). Therefore I_A=$B(p_A)$ and I_B=$B(p_B)$ follow a Bernoulli distribution with probability of success p_A and p_B respectively. Moreover

$$Y_A = \sum_{e \in P_A} I_A \approx B(n_A, p_A) \text{ and } Y_B = \sum_{e \in P_B} I_B \approx B(n_B, p_B) \quad (3)$$

are two Binomial distributions. The variables Y_A and Y_B can be approximated to a Normal distribution as follows

$$Y_A \approx N(n_A p_A, \sqrt{n_A p_A (1-p_A)})$$

and

$$Y_B \approx N(n_B p_B, \sqrt{n_B p_B (1-p_B)}) \quad (4)$$

Therefore by (4) we have that

$$Y_A - Y_b \approx N(n_A p_A - n_B p_B, \sqrt{n_A p_A (1-p_A) + n_B p_B (1-P_B)} \quad (5)$$

We are interested in construct a test to detect when the valuation of the item I is affected by the chosen scale. In order to construct the test, which is the aim of this paper, we consider the following null hypothesis:

H_0: the scale does not affect the evaluation of the item (6)

against any other alternative.

Under the null H_0 it follows that the distribution of the valuations in both scales are identical and therefore Me_A = Me_B. Therefore under H_0 we have that I_A and I_B follow a Bernoulli distribution with probability of success p_A= ½ and p_B= ½ respectively.

Thus we can restate the null H_0 in the following terms

H'_0: p_A= p_B= ½ (7)

Then under H'_0 and by (5) we can construct the following statistic which follows a standard Normal distribution:

Table 5
H-test

Scale type	1-5	–3 +3	No Uncertainty 1-7	Visual (Computer)	Visual (Paper)
Free	2.108*	–3.000*	1.664	2.604*	2.286*
Free uncertainty	2.000*	–2.393*	1.549	2.500*	2.176*
Scale type	1-5	–3 +3	Uncertainty 1-7	Visual (Computer)	Visual (Paper)
Free	1.692	–2.176*	1.838	2.356*	2.176*
Free uncertainty	1.575	–2.840*	1.721	2.255*	2.065*

* Significant results. Critical value: ±1.96

$$H = \frac{Y_A - Y_B - \left(\frac{n_A}{2} - \frac{n_B}{2}\right)}{\sqrt{\frac{1}{4}(n_A + n_B)}} \sim \mathrm{N}(0,1) \qquad (8)$$

Let α be a real number with $0 \leq \alpha \leq 1$. Let $z_{\alpha/2}$ be such that

$P(N(0,1) > z_{\alpha/2}) = \alpha/2.$

Then the decision rule for the H -test at a $100(1 - \alpha)\%$ confidence level is:

$$\begin{array}{ll} \text{If } -z_{\alpha/2} \leq H \leq z_{\alpha/2} & \text{Accept } H'_0 \\ \text{Otherwise} & \text{Reject } H'_0 \end{array}$$

Results of the H-test[1] are shown in the Table 5. Without considering uncertainty in rating scales, only the 1-7 scale is invariant. However, when uncertainty is considered, results importantly change, because the 1-5 scale also becomes invariant. The behaviour of both seven points rating scales do not change adding uncertainty, therefore, the 1-7 point scale remains invariant. Regarding the visual scales, scale invariance does not hold in any case.

Discussion, limitations and further research

In this research we have introduced a new test for analyzing scale invariance in between group designs. In addition, we have proposed a procedure to deal with uncertainty in individual responses. These two contributions complement the Martínez and Ruiz's (2011) work on measurement invariance in rating scales, because D-test is applied to analyze within group data, and does not take uncertainty into account. Therefore, our new test overcomes the possible problem arising from violation of principle of conditional independence when individuals respond to the same question several times using disparate rating scales. Also, the H-test is easy to compute and, as a nonparametric test, does not require any a priori distribution of the data nor conditions on the variances of the distributions to be tested.

This research has important implications for measuring service quality in sport consumer research, because we have expanded the work of Martínez et al. (2010), giving robustness to some conclusions and opening a door to new questions.

First of all, we find that participants prefer to evaluate service quality using scales ranging from 0 to 10 or from 1 to 10, in line with previous findings. This result reinforces the importance of analyzing scale invariance when the third-person approach is taken as a research perspective, i.e. when respondents have to answer a question using a scale proposed by researcher which differs from the scale they prefer. Again, it seems that in the context of consumer attitudes, the 10 point reference system is the preferred for individuals, at least for the Spanish culture, where the 10 points scales are constantly used to rate alumni, professors, to valuate performance, etc. It would be necessary that further research replicate this finding in other cultures, in order to ascertain if this principle is applicable to other contexts.

Second, the three considered rating scales are not preferred by individuals in their free choice of the scale to judge their quality perceptions. This implies that when one of these widely used scales is presented, invariance analysis should be carried on. In the context of service quality

[1] The *Mathematica 6.0* code written for doing the invariance analysis is available from authors upon request.

perceptions of sports services, this research finds that only the 1-7 scale is invariant. This is partially consistent with Martínez et al. (2010) results, because these authors also find that invariance holds also for the -3+3 scale. Therefore, the 1-7 scale is invariant using disparate assumptions, procedures and samples, which reinforce the conclusion that this is a proper scale to measure service quality.

Third, considering uncertainty in consumer responses can modify invariance analysis. This is an important finding, because the 1-7 scale performs equally well when uncertainty is added. However, the 1-5 scale is invariant under this condition, what would change one of the implications of Martínez et al. (2010) study. It is true that there is much more participants admitting uncertainty in the group of the 1-5 scale than in the group of the 1-7 scale, so this could be a reason of the different results. Therefore, further research should explore the behavior of the 1-7 scale when uncertainty covers a major percentage of responses.

Fourth, both visual scales (screen and paper) are not invariant. Uncertainty is presented in all of responses, beyond the tolerance level of the used instrument (digital and normal pen). This result may be caused by to possible reasons: (1) the instruments interact with the responses, i.e. the technical features of the instruments make respondents to be uncertain in their responses; or (2) participants truly reflect uncertainty in their responses. This distinction is very important because only around a 10% of participants acknowledged that their responses were uncertainty for the rating and the free scales conditions. Therefore, it seems more a method effect than a truly effect. However, we consider interesting to deep into this topic in future studies, with the aim to make sure than the vast majority of individuals are not uncertain about their judgments. Recall that Martínez et al. (2010) found a similar result; only a slight percentage of participants admitted they were uncertainty. Nevertheless, maybe using depth interviews and other qualitative studies would be interesting to guarantee that uncertainty is not a common feature of individual judgments.

Obviously, the test we propose together the importance of analyzing measurement invariance considering uncertainty must be expanded to other fields of consumer psychology, especially when self-report measures are a matter of interest. This test is easy to implement, new and robust method to analyze measurement invariance. However, one caveat other studies should address is the sample size issue. We recognize the small sample sizes of the different groups composing our study, so further research must increase sample sizes in order to increase power to detect significant departures of invariance.

In sum, this research has advanced in the consumer psychology field, where measuring attitudes of consumers is essential to management. We have proposed a procedure for testing invariance under uncertainty, based on non-parametric test and the use of triangular distribution. In line with Martínez et al. (2010) recommendation, we also advocate for using the first-person approach for measurement whence possible. Nevertheless, if the third-person approach is achieved, then our research shows the 1-7 scale should be used in order to maximize utility of the collected information, and minimize bias.

References

Amigo, S., Caselles, A., & Micó, J C. (2010). General factor of personality questionnaire (GFPQ): only one factor to understand personality? *The Spanish Journal of Psychology, 13*, 5–17.

Cohen, P., Cohen, J., Aiken, L., & West, S. (1999). The problem of units and the circumstance for POMP. *Multivariate Behavioral Research, 34*, 315–346. http://dx.doi.org/10.1207/S15327906MBR3403_2

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

Cresswell, S. L., & Eklund, R. C. (2006). The convergent and discriminant validity of burnout measures in sport: A multi-trait/multi-method analysis. *Journal of Sports Sciences, 24*, 209–220. http://dx.doi.org/10.1080/02640410500131431

Gardner, D. G., Cummings, L. L., .Dunham, R. B., & Pierce, Jon L. (1998). Single-item versus multiple item measurement: An empirical comparison. *Educational and Psychological Measurement, 58*, 898–915. http://dx.doi.org/10.1177/0013164498058006003

Gregg, M. & Hall, C. (2006). Measurement of motivational imagery abilities in sport. *Journal of Sport Sciences, 24,* 961–971. http://dx.doi.org/10.1080/02640410500386167

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

Hersh, H. M., & Caramazza, A. (1976). A fuzzy set approach to modifiers and vagueness in natural language. J*ournal of Experimental Psychology, 105*, 254–276. http://dx.doi.org/10.1037//0096-3445.105.3.254

Kilpatrick, F. P., & Cantril, H. (1960). Self-anchoring scaling: A measure of individuals' unique reality worlds. *Journal of Individual Psychology, 16,* 158–173.

Ko, Y. J., & Pastore, D. L. (2005). A hierarchical model of service quality for the recreational sport industry. *Sport Marketing Quarterly, 14*(2), 84–97.

Martindale, R. J. J., Collins, D., Wang, J. C. K., McNeill, M., Lee, K. S., Sproule, J., & Westbury, T. (2010). Development of the talent development environment questionnaire for sport. *Journal of Sports Sciences, 28*, 1209–1221. http://dx.doi.org/10.1080/02640414.2010.495993

Martínez, J. A., Ko., Y. J., &. Martínez, L. (2010). An application of fuzzy logic to service quality research: A case of fitness service. *Journal of Sport Management, 24*, 502–523.

Martínez, J. A., & Ruiz, M. (2011). D-test: A new test for analyzing scale invariance using symbolic dynamics and symbolic entropy.

Methodology. 7, 88–95. http://dx.doi.org/10.1027/1614-2241/a000026

Nicolai, A. T., & Dautwiz, J. M. (2010). Fuzziness in action: What consequence has the linguistic ambiguity of the core competence concept for organizational usage? *British Journal of Management, 21*, 874–888. http://dx.doi.org/10.1111/j.1467-8551.2009.00662.x

San Martín, J., Perles, F., & Canto J. (2010). Life satisfaction and perception of happiness among university students. *Spanish Journal of Psychology, 13*, 617–628.

Scherpenzeel, A. C., & Saris, W. E. (1997). The validity and reliability of survey questions: A meta-analysis of MTMM studies. *Sociological Methods and Research, 25*, 341–383. http://dx.doi.org/10.1177/0049124197025003004

Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods, 1*, 199–223. http://dx.doi.org/10.1037//1082-989X.1.2.199

Schmidt, F. L., & Hunter, J. E. (1999). Theory testing and measurement error. *Intelligence, 27*(3), 183–198. http://dx.doi.org/10.1016/S0160-2896(99)00024-0

Shamir, B., & Kark, R. (2004). A single-item graphic scale for the measurement of organizational identification. *Journal of Occupational and Organizational Psychology, 77*, 115–123. http://dx.doi.org/10.1348/096317904322915946

Spanos, A. (2010). The discovery of Argon: A case for learning from data? *Philosophy of Science, 77,* 359–380. http://dx.doi.org/10.1086/652961

Zaltman, G., & Zaltman. L. (2008). *Marketing metaphoria: What deep metaphors teveal about the minds of consumers.* Boston, MA: Harvard Business School Press.

Received March 10, 2011
Revision received August 6, 2011
Accepted September 15, 2011