Is there an Optimal Number of Scale Points to Measure Attitudes and Preferences in Online Surveys among Hispanics?

Myra Mabel Pérez Rivera
University of Puerto Rico, Río Piedras Campus

Abstract

An exploratory experimental research procedure was followed to examine the consistency of responses to different Likert type scale formats employed in online surveys among Hispanics. The purpose of the study was to determine if there is an optimal number of scale points for assessing attitudes and preferences to ensure cross-cultural comparability and measurement equivalence while conducting studies online. The participants were Hispanic undergraduate students with residence in Puerto Rico. The method followed was an adaptation of Matell and Jacoby (1971, 1972) that considered Nunnally (1978) suggestions on dealing with scale reliability measurement issues. The Likert scales used were the ones with numerical descriptors and verbal anchors. The analysis revealed that a significant difference exist in the reliability of responses to the Likert type scale formats of seven, five and three points for online studies among Hispanics in Puerto Rico. The results suggest that the Likert type scale of seven points is the most reliable and may be preferred to conduct studies online among Hispanics in Puerto Rico.

Key words – Online marketing research, Cross-cultural research, Likert scales, Reliability testing, Response quality

Literature Review

A great concern for marketing researchers is the development of equivalent and comparable instruments that can be useful in assessing attitudes and preferences across cultures. One important issue in the construction of scales is the determination of the optimal number of response categories that will let us discriminate between rated items and that will give us consistent and reliable responses. National studies related to this issue have indicated that reliability and consistency are independent of the number of scale points, but cross-cultural studies indicate that the same scale may have different reliabilities in different countries. In this regard, Parameswaran and Yaprak (1987) have

stated that the same instrument used in a cross-national survey may lead to different levels of response reliabilities among various country samples due to difference in knowledge, perceptions, familiarity with research instrument and the national propensity to certain response style.

There have been various studies on the topic within the context of traditional modes of survey administration (telephone, face to face, self-administered, and assisted), but none had evaluated online surveys as contact method. This study explores this issue by comparing the consistency of responses to different Likert type scale formats used in online studies among Hispanics.

Online Marketing Research

The exponential growth of the Internet has induced one of the most profound developments in survey methodology (Dillman, 2000), namely, the collection of survey data via the Internet. Online surveys are increasingly used in both academic (Mandel and Johnson, 2002; Meuter et al., 2000) and market research (Deutskens, Jong, Ruyter and Wetzels, 2006). Persuasive arguments for choosing online surveys over traditional methods include benefits such as lower costs, faster response times, and wider geographic reach, which makes them especially suitable for cross-national research (Dillman, 2000; Illieva, Baron and Healey,2002; Roster, Rogers, Albaum and Klein, 2004).

Proponents of online surveys argue that the Internet allows for the use of uncomplicated directions, as well as richer and more interesting question formats. Online surveys also have been found to be useful in reaching busy professionals, a population for whom mail surveys suffer from low and continually declining response rates. They can fill them in at their convenience and can partially complete and return whenever they like. Other advantages highlighted relate to how the use of new technology in online surveys allows research that is more visual, flexible and interactive.

Another advantage suggested is that online surveys do not require interviewers to be present and so interviewer effects area avoided. This is likely to be a significant advantage for certain types of study particularly where social desirability effects are likely to be large (Duffy et al 2005). It is argued that this may help explain the more 'socially liberal' attitudes seen in many online surveys, as respondents on average tend to lead less home-based lives and so are less cautious (Kellner 2004). Indeed this argument suggests that this feature could actually produce achieved samples that are more representative than traditional approaches, as online interviewing reaches 'busy people, often educated and well-off who systematically repel or ignore cold callers but are willing to answer questions posted on their computer screen' (Kellner 2004). However, others argue that it is the fact that online respondents are more 'viewpoint oriented' (i.e. more likely to have active opinions) that accounts for the different attitudes seen in online surveys.

Although extensive evidence details response rates and completion times in online surveys (Deutskens, Ruyter, Wetzels and Oosterveld, 2004; Illieva, Baron and Healey, 2002), evidence about their quality seems sparse and inconclusive (Stanton, 1998). Several authors have found, for example, that online and mail surveys produce different results (McDonald and Adam, 2003; Shermis and Lombard, 1999) that cannot be equated through a simple weighting factor or adjustment strategy (Deutskens et al 2006). Other disadvantages cited for internet-based methodologies focus mainly on sampling issues. Unlike face-to-face surveys, which can be sampled from reasonably comprehensive databases, online surveys are most often conducted among respondents from a panel who have agreed to be contacted for market research. No simple database of everyone who is online exists, and it looks unlikely to exist for the foreseeable future. Furthermore, even if there were such a list, prohibitions against 'spamming' online users would prevent it from being used as a sampling frame. There are therefore three main issues relating to coverage bias or selection error that are raised with the sampling approach to online panel surveys: first, of course, they can reach only those who are online; second, they can reach only those who agree to become part of the panel; and, third, not all those who are invited respond (Duffy et al 2005).

Other issues are raised around mode effects, where, for example, it is known that online respondents use scales differently from respondents in other modes. There is conflicting research on this, some showing that online respondents are more likely to choose midpoints in scales and 'don't know' options in general, and other research, in contrast, suggesting that online respondents tend to choose extreme responses on these scales.

Either way, these types of effect will be due to complex competing effects of response styles and do not necessarily make responses from online surveys less accurate, but they can cause problems when we attempt to switch to an online survey approach in tracking work. In this regard, Duffy et al (2005) stated that it is possible to correct for this to an extent through modeling, but this is likely to be viewed as less straightforward for those commissioning.

In summary, online research is definitely a research growth area. More than just a new technology, it represents a change in the way we conduct and think about research. With raw data just a few clicks away, most of us understand the time and cost advantages of an online medium (Miller 2001). Due to the easy and cheap access to the Internet and new software solutions, it is quite easy to conduct the data collecting over the Internet without any direct contact to the interviewer (Klein, Nihalani and Krishnan, 2010). By using the Internet as a medium to contact people, geographical barriers do not count as much as before and people can be interviewed, who could not have been interviewed before (Klein, Nihalani and Krishnan, 2010; Couper, Tourangeau & Kenyon, 2004). Nevertheless, it is questionable if the data, which is gathered over the Internet, is of the same quality as data, which is gathered through personal interviews, where quality of data refers to data that provides a reliable basis for decision making (Deutskens, Jong, Ruyter and Wetzels, 2006). Therefore, the following question arise: Are online measures comparable to traditional measures (Miller 2001). This question of comparability leads to the following questions: 1. Are online measures reliable for decision making? and 2.Are online measures comparable across cultures?

Cross-cultural Research

Cross-cultural researchers need to ensure that their concepts and methods are appropriate and valid across all cultures under investigation (Baumgartner and Steenkamp 2001, Durvasula, Andrews, Lyonski and Netemeyer, 1993; Gleason, Devlin, and Brown 2003; Steenkamp and Baumgartner 1998; Van de Vijver and Leung 1997). This can be a formidable task, as Americans have designed the vast majority of these concepts and measures and may not necessarily be cross-culturally applicable (Gorn 1997; Wong, Rindfleisch, Burroughs and Steenkamp, 2003).

Measurement equivalence - One of the major challenges in marketing research has been to ensure the equivalence of measurements across a sample of respondents. This is particularly true in cross-national research, because each country is characterized by a unique pattern of languages, values, and socio-cultural behaviors. As a result, international marketing researchers must approach measurement issues with caution and realize that measurement inequivalence across unique country populations can detrimentally affect the value of the research effort (Myers, Clantone, Page, and Taylor, 2000).

A good scale must, first and foremost, stand the test of reliability and validity. In addition, the choice of scale, number of scale categories or points, and anchor words should be such that the instrument does not generate response bias, and produces actionable or managerially relevant results (Agarwal 2003). Recent work has provided

consumer researchers with a number of tools and techniques for assessing and improving cross-cultural validity and reliability of consumer behavior scales (Chang, 1994,: Davis, Douglas and Silk 1981; Wong et al 2003).

Comparability is an important issue in cross-cultural research. Hence, a great concern for marketers is the development of equivalent and compatible instruments that will be useful in assessing consumer attitudes and preference across cultures. To ensure comparability of behavior across cultures each culture should be understood in its own terms (Pareek and Rao 1980). Each culture has their own way of explaining experience, traditions, customs and norms. Therefore, the question asked to their members, the words use in questions should appear natural to the particular setting. However, in order to compare findings across cultures, the information should have to be equivalent (Bhalla and Lin 1987). This means that the instruments have to be made cross-culturally equivalent (Guttman, 1967, Pareek and Rao 1980).

As stated by De Beuckelear, Lievens and Swinnen (2007) cultural values might also influence how individuals interpret the rating scale (Riordan and Vandenberg, 1994). Specifically, prior cross-cultural research has shown that the differences between the intervals of a rating scale are differently perceived across cultures. In fact, substantial cross-country differences have been found with regard to the tendency to agree with items, regardless of the item content (Riordan and Vandenberg, 1994; Ross and Mirowski, 1984; Van Herk, Poortinga, and Verhallen, 2004). Similarly, there is empirical evidence for cross-country bias due to the respondents' use of extreme

responses on rating scales as this bias exists between Korean and American respondents, Japanese and American respondents, and French and Australian respondents. Such cross-country differences in response styles produce systematic differences in observed variable means and variances (De Beuckelear et al 2007). As a result, the assumption of measurement equivalence of survey instruments may not be tenable. Although these prior studies were not conducted in an organizational (survey) context, they might have direct implications for organizational surveys because the latter also use rating scales.

Some recent research (Liu, Borg, and Spector, 2004; Ryan, Chan, Ployhart and Slade, 1999) has examined similar questions with regard to the cross-cultural equivalence of organizational surveys across multiple countries. Ryan et al (1999) scrutinized the equivalence of an organizational survey of a multinational company across four countries (Mexico, U.S., Australia and Spain). They found that the organizational survey was equivalent across U.S. and Australian samples only. Recently, Liu et al (2004) examined whether the German Job Satisfaction Survey was 'transposable' across 18 countries. These countries were located in four cultural regions of Schwarz's (1999) cultural model, namely West Europe, Far East, English-speaking region and South America (i.e. Latin cultural region). Two other regions (East Europe and Islamic countries) were not included in their study. Liu et al (2004) concluded that the German Job Satisfaction Survey was equivalent only across countries sharing the same cultural background and language. For example, measurement equivalence was established across countries within the same cultural region. In addition, the satisfaction survey was

more equivalent among countries in similar cultural regions than among countries in distant cultural regions.

When psychological and work-related constructs are measured in a cross-cultural context, it is pivotal to establish equivalence of the measures prior to drawing meaningful substantive conclusions about the relative importance of constructs across countries (Riordan and Vandenberg, 1994; Schaffer and Riordan, 2003; Vandenberg and Lance, 2000; Van de Vijver and Leung, 1997). Lack of measurement equivalence in data across countries implies that there is no common basis to compare data across countries: In such case, observed mean differences on relevant constructs (across countries) might result from measurement artifacts related to the measurement instrument used rather than from true differences across countries.

Establishing measurement equivalence enables us to answer a series of important questions (see Table 1) such as: Do respondents in different countries use a similar frame-of-reference when answering items used to measure relevant constructs? Do respondents in different countries calibrate the intervals on the measurement scale used in similar ways? Are differences in response styles across countries (the tendency to say 'yes' or to use extreme response categories) partly responsible for observed cross-country differences in mean item scores?

Table 1. Measurement equivalence tests and their conceptual meanings

Measurement equivalence model	Statistical test	Conceptual meaning	Implications
Form Equivalence model	Equivalent pattern of salient and non-salient factor loadings across countries.1 To set a metric for the factor, the loading of one indicator per factor (i.e. the reference indicator) is constrained to one in all countries	There are no cross-country differences in respondents' frame-of-reference when completing the instrument	All factors are measured by an identical set of indicators in all countries
Metric Equivalence model	All factor loadings are constrained to be identical across countries	There are no cross-country differences in respondents' calibration of the intervals on the measurement scale. Differences in Extreme Response Style (ERS) across countries are not likely.	Structure-level comparisons (i.e. dealing with cause-effect relationships) across countries are meaningful
Scalar Equivalence model	All factor loadings and indicator intercepts are constrained to be identical across countries	Differences in Acquiescence Response Style (ARS) (i.e. agreement bias) across countries are not likely	Structure-level comparisons and level oriented comparisons (based on estimated construct means) across countries are meaningful

Source: De Beuckelaer, Lievens and Swinnen 2007

Scalar equivalence - Cross-cultural equivalence requires linguistic equivalence, construct equivalence and scalar equivalence. Of these, scalar equivalence is the one related to comparability of scales and scaling procedures. Scalar equivalence is achieved when two individuals from separate cultures with the same value on the same hypothesized variable will score on the same level on the same test (Bhalla and Lin 1987). Scalar equivalence is very difficult to attain because cultures differ in their response set characteristics, such as social desirability, acquiescence, and evasiveness, which influence response scores (Bhalla and Lin, 1987; Toyne and Walters, 1993; Van de Vijver and Poortinga 1982).

With respect to scalar equivalence, Craig and Douglas (2001) have indicated that two aspects have to be considered in determining scalar equivalence. The first concerns the specific scale or scoring procedure used to establish the measure and whether relationships among these are patterned similarly in different contexts, and the equivalence of responses to a given measure in different countries. The greater the emphasis placed on quantitative measurement and data interpretation, the more important the establishment of scalar equivalence becomes. Scalar equivalence in scale and scoring procedures is of particular relevance insofar as the most graduation of scales or scoring procedures may vary from one country to another. A second aspect of scalar equivalence concerns the response to a score obtained on a measure. The question is whether a score obtained in one research has the same meaning and interpretation in another (Douglas and Craig, 1990).

The use of identical procedures in different cultures for eliciting and quantifying data does not necessarily result in the measurement of the same variable since the manifest response may have different meanings in different cultures, and similar phenomena may be categorized differently based upon meanings and applications in context of their individual norms and values (Choudry 1986; Strauss 1969). Therefore, to ensure comparability across cultures it is imperative that scales be tailor-made and/or carefully tested in each culture in terms of relevance and appropriateness (Davis, Douglas and Silk, 1981; Onkvisit and Shaw 1989; Singh 1995).

Likert scales

Rating scales are one of the most widely used tools in marketing research and commercial market research. They are used to capture information on a range of phenomena. In consumer research, respondents may be asked about their attitudes, perceptions or evaluations of products, brands or messages – among many other possibilities (Dawes 2008).

Rating scales typically require the respondent to select their answer from a range of verbal statements or numbers. The Likert scale is a research tool widely use in marketing research to measure attitudes and preferences. Rensis Likert introduced this method of summated ratings in 1932. The original Likert method consisted of asking subjects to respond to attitude statements on a 5-point response scale containing labels of STRONGLY DISAPROVE, DISAPROVE, UNCERTAIN, APROVE, and STRONGLY APROVE. Further developments of the scale made use of a variety of scale labels and made use of different numbers of point scales (Likert 1967; Wyatt and Meyers 1987). An example of the Likert response scale is as follows: strongly disagree, disagree, neither disagree nor agree, agree, strongly agree. This particular example is a 5-point Likert scale utilizing verbal response descriptors. Likert scales may also use numerical descriptors where the respondent selects an appropriate number to denote their level of agreement. For example, a question could be worded like this: 'Indicate your agreement from 1 to 5 where 1 equals strongly disagree and 5 equals strongly agree' (Dawes 2008).The range of possible responses for a scale can vary. Textbooks on the subject

typically portray 5- or 7-point formats as the most common (Malhotra & Peterson 2006, ch. 10); 10- or 11-point scales are also frequently used (Loken, Pirie, Virnig 1987).

Two types of analysis are commonly carried out on sets of Likert responses. The first type relates to score building. Responses to items are treated as belonging to a numerical scale, and are either summed over the items, or a factor or latent variable analysis is carried out, and a weighted or unweighted score is produced, which is taken to measure a common characteristic of the item set for a respondent (Dittrich, Francis, Hatzinger and Katzenbeisser 2007). The second type of analysis is concerned more with providing an ordering of the relative importance of a set of items, and how this relative importance might vary according to other characteristics of the individual (Dittrich et al 2007).

Commonly, simple methods are used to examine the relative importance of Likert Items (Dittrich et al 2007). Sometimes, Likert items are treated as categorical. Other studies look at the percentage of responses in a particular combination of responses. Another common procedure is to treat each Likert scale as continuous: In such cases a mean and standard deviation is often reported for each of the Likert-scale questions and the items are ranked according to the means. The effect of subject covariates for each item separately can also be investigated to account for differences between groups. More sophisticated methods might use a multivariate approach, and simultaneously analyze the joint pattern of means for the set of items through MANOVA or multivariate regression. These approaches can be problematic for a variety of reasons. Simple categorical approaches either fail to utilize the complete information in the data, or have difficulty in determining a proper ranking of the items. In addition, much analysis is

descriptive and lacks proper statistical analysis when comparing groups. Methods analyzing means (either univariate or multivariate) assume both that the distance between response categories are equal, and that the responses have an underlying normal or multivariate normal distribution. These assumptions made are often unrealistic in practice (Dittrich et al 2007).

Other classes of methods rely on models based on latent variable approaches (Dittrich et al 2007). The ordered categorical scale is assumed to be a manifestation of a latent quantitative variable and could be analyzed by the proportional odds model. However, a common problem with all these methods is that Likert responses are treated as *absolute* measurements, and this can be a rather dubious assumption especially when dealing with subjective self-assessments. In the psychometric literature it is a basic assumption that one requirement for defining measurements is that individuals giving the same answer to a Likert item (choosing the same category) do not only share the same response value but are equivalent with respect to the attitudes, values, etc. to be measured. Brady (1989) addresses this problem in the context of factor and ideal point analysis for interpersonally incomparable data. This is a particular problem when comparing different countries or cultures (Heine, Lehman, Peng, and Greenholz 2002).

Reliability of scales

With more advertisers and marketing researchers turning to online research, the reliability of the scales used for collecting the data has taken on great importance. Unreliable measuring instruments can result in distortions and instability of the

segmentation structure, possibly leading to incorrect marketing decisions and defective strategies. The basic notion underlying reliability is consistency: Does the questionnaire item yield the same answer from a given individual when that person responds to the item on several occasions (Boote 1981)? By definition, a durable state of mind will be stable over relatively long intervals of time. Thus, questionnaire items that are not reliable cannot be valid measures of the respondents' values or other durable states of mind. This research issue must be addressed within the context of management's information requirements and management needs to have confidence in the information provided. While acknowledging the importance of reliability, little empirical work has been reported. Research concerned with reliability of Likert-type attitude-scale items has concentrated primarily on the issue of the number of scale points (Bendig, 1954; Burns and Harrison, 1979; Jacoby and Matell, 1971; Komorita and Graham, 1965; Lehman and Hulbert, 1972). National studies related to this issue are contradictory. For example, Jacoby and Matell (1971) have asserted that reliability is independent of the number of scale points. Conversely, Bendig (1954) has argued that a scale's reliability and the number of its scale points are related. The evidence to support these arguments is by no means unequivocal. To some extent, the relationship between reliability and the number of scale points is affected by the actual measure of reliability applied to the data. For example, Lehman and Britney (1977) point out that the correlation coefficient in a test/retest situation will increase with the number of scale points while the proportion of completely consistent responses is inversely related to the number of scale points. The effects of labeling or anchoring scale points have been given less attention by researchers. Bendig (1953), in an early study, found that "increased verbal definition of

the categories resulted in slightly increased reliability." In addition, some work has been done on the effects attributable to the use of different measurement techniques and various statistical tests of reliability (Lehman and Britney, 1977), and how a scale's content will affect its reliability (Burns and Harrison, 1979; Komorita and Graham, 1965), as well as the relative reliability of scale ratings versus item ranking (Munson and Mc-Intyre, 1979; Reynolds and Jolly, 1980).

On the other hand, cross-cultural studies suggest that the same scale may have different reliabilities in different countries (Davis, Douglas and Silk, 1981). Parameswaran and Yaprak (1987) stated that the same research instrument used in a cross-national survey may lead to different levels of response reliabilities among various country samples due to difference in knowledge, perceptions, familiarity with research instrument and the national propensity for certain response styles (Davis, Douglas and Silk 1981; Parameswaran and Yaprak 1987). In this regard, Onkvisit and Shaw (1989) have stated that even though Likert scales have proven to be satisfactory in measuring behavior and opinion in the United States, they may not elicit the same manner of response in other markets. A seven point scale, for example may not yield more information than a five point scale in the United States, but the former may prove useful elsewhere (Flaskerud 1988; Onkvisit and Shaw 1989; Lee, Jones, Mireyama and Zhang 2002; Heine, Lehman, Peng and Greenholtz 2002). Moreover, it is suggested that difficulties with Likert scale formats may be due to such factors as education or faulty translation, but it is also possible that the degree of variation that Likert scales attempt to measure is meaningless to some cultural groups (Flaskerud 1988).

With respect to the Hispanic market, Stanton, Chandran and Hernandez (1982) have suggested that scales of the three point variety should be used in Latin America because respondents in developing countries with low literacy levels may not be familiar with very fine shades of the meanings of the questions asked. As an example, in a study conducted in Mexico City it was observed that subjects ignored the response format given in the questionnaire and instead expressed their own opinions in their own words (Pick de Weiss and Jones 1981). In this regard, Hernández and Kaufman (1990) recommended the usage of simplified scales; specifically they suggested the truncation of the Likert scale into a 3-point scale because this will facilitate understanding. In addition, a comparative study between Hispanics and non-Hispanics in the United States revealed that reliability of responses to Likert type scales is affected by level of acculturation (Pérez-Rivera 1994). An alternative scale that may provide a consistent and reliable response among Hispanics is the graphic scales of faces (Pérez-Rivera 1994, 1996). The graphic scales are easy to answer because the respondent marks a pictorial depiction of his level of agreement, but a disadvantage of the graphic scale is that it may be disturbing for those respondents who are familiar with Likert type scale formats (Worchester and Downhan 1991).

Research Objectives and Methodology

The purpose of this exploratory experimental study was to examine the consistency of responses (test-retest reliability) to Likert type scale formats employed in online surveys

in order to determine if there is an optimal number of points or scale formats for assessing attitudes and preferences among Hispanics to ensure cross-cultural comparability, specifically scalar equivalence. The research questions were:

1. Is the consistency (test-retest reliability) of responses in online research independent of number of scale points?
2. Is there an optimal scale format for Hispanics in online research?

Scale reliability - Scale reliability consists of two different components: stability and equivalency. The former represents temporal stability of a measure at two different points in time, while the latter is more focused on the internal consistency or internal homogeneity of the set of items forming the scale. This study focuses on the temporal stability by using a test-retest method as a way to examine the reliability of the scale format used. The rational for using a test-retest approach is that first, the scale used in the study as defined as an attitudinal judgment, is assumed to be stable within a short period of time. Secondly, adequate sample size should be able to even out the true changes in attitude. Nunnally (1978) suggest that, if the instrument is indeed stable, the two administrations should have a resulting correlation of .80 or higher.

Three potential problems must be addressed when applying the test retest methodology: recall, time and reactivity (Nunnally, 1978). A recall problem may arise when subjects are administered the instrument again within too short an interval. Subjects may recall their first responses at the second administration of the measure. Similarly, a time problem may arise if subjects are tested within too long an interval. True changes in the

subjects concerned, rather than inconsistencies in the instrument, leads to a different test-retest result. Lastly, a problem with reactivity can occur when subjects are administered the instrument multiple times since they become sensitized to the instrument and learn to answer as they are perceive they are expected to respond. Nunnally (1978) recommends that a period of two weeks to one month should be elapse between test and retest administrations in order to minimize the memory effect (Nunnally and Bernstein 1994; Lam and Woo 1997).

Procedure - The procedure followed to conduct the study was similar to the one used by Jacoby and Matell (1971 and 1972). The study was conducted in two phases, with three week interval time between phases. Participants were contacted through emails and they had to answer an online questionnaire that measured attitudes towards war. To match the questionnaires the respondents were asked to identify themselves in both questionnaires with a code of 4 to eight letters or numbers only known by them. The scale statements were back translated from English to Spanish and then back to English to ensure its meaning by a certified translator. The scale used was extracted from the book Scales for the Measurement of Attitudes by Shaw and Wright (1967). Participants were randomly assigned to a different scale format: Likert of 3 points, Likert of 5 points and Likert of 7 points. The participants were all students from the University of Puerto Rico, Rio Piedras campus.

The researcher considered that the topic attitude towards war elicits fairly stable attitudes that do not change during the three week interval between phases. In addition,

the researcher considers that the three week interval time is long enough to minimize the effect of memory and short enough to minimize the effect of a change in attitude of respondents due to other variables.

The dependent variable of the study was the consistency of responses to different scale formats expressed as the test-retest reliability coefficient. The independent variables were the scale formats: Likert of 3 points, Likert of 5 points and Likert of 7 points. Specifically, the Likert scale format used were numerical descriptors with verbal anchors.

Method of Analysis - The responses obtained from the test and retest reliability were analyzed to determine the test-retest reliability coefficients for each participant in each scale format. Then, a fisher transformation was used to convert all reliability coefficients to normality. These transformations were then analyze by a simple analysis of variance and through MANOVA repeated measures.

Research Findings

A total of 225 questionnaires were sent twice via email, of these 84 questionnaires where completed for phase 1 and 75 for phase 2. However, 45 questionnaires were adequately completed and could be matched for a net response rate of 20%.

The analysis revealed that a significant difference exist in the reliability of responses to the Likert type scale formats of seven, five and three points for online studies among Hispanics in Puerto Rico (see Table 2). The results suggest that the Likert type scale of seven points is the most reliable and may be preferred to conduct studies online among Hispanics in Puerto Rico.

Table 2 – ANOVA for Test-retest Reliability Coefficients by Likert-type Scale Format

Likert- type Scale Format	Test-retest reliability coefficients	N	F ratio and significance value
Likert 3 points	.5676	15	F= 4.713
Likert 5 points	.6118	15	p= .014
Likert 7 points	.7494	15	

Discussion

A far as can be determined this study is the first attempt to address the relationship between reliability of responses with respect to type of scale format and number of points among Hispanics in online studies.

A growing number of consumer researchers have embarked on cross-cultural research in order to understand, explain and predict behavior of participants of our global consumer culture. Unfortunately, the generalizability of the conceptualizations and measures employed in these studies remain unclear, as few studies have examined their cross-cultural applicability. Thus, the inquiry into cross-cultural measurement equivalence,

comparability, and reliability of scales provides an important and much needed foundation for future cross-cultural research. This study enable researchers with an understanding on how to evaluate measurement reliability across samples to ensure comparability and to disentangle cultural differences in instrument usage from measurement related differences while taking advantage of online technology.

In summary, the results indicate that the reliability of responses is affected by number of scale points among Hispanics in online studies. The results suggest that the Likert scale of 7 points should be preferred among Hispanics to conduct studies online.

Threats to the external validity of the study are the sample selected and the number of scale points used. Therefore, further research should be conducted to determine whether the present findings can be generalized to a different population (non students) defined by such parameters as level of education or ability, socio-demographic characteristics and beyond:

1. Hispanics in Puerto Rico (e.g. Hispanics in the United States, Costa Rica, Panama)
2. The number of scale points (e.g. 4, 6, 8 and 10 points)
3. The Likert -type scale format (e.g. Diagrammatic scale, Semantic differential))

References

Agarwal, S. 2003, "The Art of Scale Development," *Marketing Research*, 15(3) 10-13.

Baumgartner, H. and J.E.M Steenkemap 2001 "Response Styles in Marketing Research: a Cross- national Investigation," *Journal of Marketing Research*, 28(May), 143-156.

Bendig, LA.W. 1954. "Reliability and The Number of Rating Scale Categories. *"The Journal of Applied Psychology* 38 (1): 38-40.

Bhalla, G. and Lin. L.Y.S. 1987. "Cross-Cultural Marketing Research: A Discussion of Equivalence Issues and Measurement Strategies". *Psychology and Marketing* 4(4):275/285.

Boote, A.S. 1981. "Reliability Testing of Psychographic Scales". *Journal of Advertising Research* 21(5):53-60.

Brady, H.E. 1989. "Factor and Ideal Point Point Analysis for Interpersonably Incomparable Data," *Psychomerika*, 54, 181-202.

Burns, Alvin C, and Harrison, M.C. 1979. "A Test of Reliability of Psychographics,"\ *Journal of Marketing Research*, 16(1979):32-38.

Chang, L. 1994, "A Psychometric Evaluation of 4-point and 6-point and 6-point Likert-type Scales in Relation to Reliability and Validity", *Applied Psychological Measurement*, 18(3), 205-215.

Choudry, Y.A. 1986. "Pitfalls in International Marketing Research; Are you speaking French Like a Spanish Cow?" *Akron business and Economics Review* (Winter): 18-28.

Couper, M., Tourangeau, R., and Kenyon, K. 2004. "Picture This! Exploring Visual Effects in Web Surveys," *Public Opinion Research*, 68, 255-266.

Craig, C.S. and Douglas, S.P. 2001, Conducting International Market Research in the Twenty-first Century, *International Marketing Review*, 18(1), 80-90.

Davis, H.L., Douglas S.P. and Silk, A.J. 1981, "Measure Unreliability: a Hidden Threat to Cross-National Marketing Research," *Journal of Market*ing, 45(2), 98-109.

Dawes, J. 2008, "Do Data Characteristics Change According to the Number of Scale Points Used?" *International Journal of Market Research*, 50(1) 61-77.

De Beuckelaer, A., Lievens, F. and Swinnen, G. 2007, "Measurement Equivalence in the Conduct of a Global Organizational Survey across Countries in Six Cultural Regions," *Journal f Occupational and Organizational Psychology*, 80, 575-600.

Deutskens, E., de Jong, A., de Ruyter, K. and Wetzels, M. 2006, "Comparing the Generalizability of Online and Mail Surveys in Cross-National Service Quality Research," *Marketing Letters*, 17:119-136

Deutskens, E.C., de Ruyter, K., Wetzels, M.G.M., and Oosterveld, P. (2004). Response rate and response quality of internet-based surveys: An experimental study. *Marketing Letters, 15*(1), 21–36.

Dillman, D.A. (2000). Mail and internet surveys. *The tailored design method*, New York: Wiley.

Dittrich, R., Francis, B., Hatzinger, R. and Katzenbeisser, W. 2007, "A Paired Comparison Approach for the Analysis of Sets of Likert-scale Responses," *Statistical Modelling,* 7(1), 3-28.

Duffy, B., Smith, K., Terhanian, G. and Bremer, J. 2005, "Comparing Data from Online and Face-to-face Surveys," *International Journal of Market Research,* 47(6), 615-639.

Durvasula, S., Andrews, J.C., Lyonski, S. and Netemeyer, R. 1993, "assessing Cross-National Applicability of Consumer Behavior Models" *Journal of Consumer Research,* 19(March), 626-636.

Flaskerud, J.H. 1988, "Is the Likert Scale format culturally biased?" *Nursing Rese*arch, 37(3), 185-186.

Gleason, T.C., Devlin, S.J. and Brown, M. 2003, *"In Search of the Optimum Scale*," Marketing Research, 15(3) 25-30.

Gorn, G., 1997, "Breaking out of the North American Box," in Advances in Consumer Research, 24, 6-8.

Guttman, L. 1967. "A Basis for Scaling Data." In Fishbein, M. *Attitude Theory and Measurement*. New York: John Wiley and Sons. 96-107.

Heine, S.J., Lehman, D.R., Peng, K. and Greenlholtz, J. 2002, "What's Wrong With Cross-Cultural Comparisons of Subjective Likert Scales?" *Journal of Personality and Social Psychology,* 82(6), 903-919.

Illieva, J., Baron, S., and Healey, N.M. 2002 "Online surveys in marketing research: Pros and cons," *International Journal of Market Research, 44*(3), 361–382.

Jacoby, J and Matell, M.S. 1971. "Three-Point Likert Scales are Good Enough." *Journal of Marketing Research*. 8(November): 495-500.

Kellner, P. 2004, "Can Online Polls Produce Accurate Findings?" *International Journal of Market Research,* 46, 1.

Klein, A., Nihalani, K., and Krishnan, K.S. 2010, "A Comparison of the Validity of Interviewer-based and Online-conjoint Analyses," *Journal of Management and Marketing Research,* 2(1) 1-15.

Komorita, S.S. and Graham, W.K. 1963. "Number of Scale Points and the Reliability of Scales". *Educational and Psychological Measurement* 25(4): 987-995.

Lee, J.W., Jones, P.S., Mineyama, Y. and Zang, X.E., 002, "Cultural Differences in responses to Likert scale," Research in Nursing Health, 25(4), 295-306.

Lehman, Donald R, and Britney, Kathryn E. A. 1977, "Determining an Appropriate Measure of Reliability for Psychographic Measure," In *Contemporary Marketing Thought,* ed, Barnett A. Greenbert and Danny N. Bellenger. Chicago: American Marketing Association.

Lehman, D. R., and Hulbert, J. 1972, "Are Three-Point Scales Always Good Enough?" *Journal of Marketing Research 9*:444-46.

Likert, R. 1967. "The Method of Constructing an attitude Scale." in Fishbein, M. *Attitude Theory and Measurement*. New York: John Wiley and Sons. 90-95.

Liu, C., Borg, I., and Spector, P. E. 2004, "Measurement equivalence of the German job satisfaction survey used in a multinational organization: Implications of Schwartz's culture model." *Journal of Applied Psychology*, 89, 1070–1082.

Loken, B., Pirie, P., Virnig, K. *et al*. 1987, "The use of 0–10 scales in telephone Surveys," *Journal of the Market Research Society,* 29, 3, July, pp. 353–362.

Malhotra, N. & Peterson, M. 2006 *Basic Marketing Research: A Decision-Making Approach* (2nd ed.). New Jersey: Prentice Hall.

Mandel, N. and Johnson, E.J. 2002. "When web pages influence choice: Effects of visual primes on experts and novices," *Journal of Consumer Research, 29*(2), 235–245.

Matell, M.S. and Jacoby, J. 1971, "Is there an Optimal Number of Alternatives for Likert Scale items? Study I: Reliability and Validity." Educational and Psychological Measurement 31: 657-674.

Matell, M.S. and Jacoby, J. 1972. "Is there and Optimal Number of Alternatives for Likert-Scale Items?" *Journal of Applied Psychology* 56(6): 506-509.

McDonald, H., & Adam, S. (2003). A comparison of online and postal data collection methods in marketing research. *Marketing Intelligence & Planning, 21*(2), 85–95.

Meuter, M.L., Ostrom, A.L., Roundtree, R.I., and Bitner, M.J. 2000. "Self-service technologies: Understanding customer satisfaction with technology-based service encounters," *Journal of Marketing, 64*(3), 50–64.

Miller, T.W. 2001, "Can we Trust the Data of Online Research," *Marketing Research*, 13(2) 26-32.

Munson, J. M., and Mclntyre, S. H., 1979, "Developing Practical Procedures for the Measurement of Personal Values in Cross-Cultural Marketing," *Journal of Marketing Research* 16:48-52,

Myers, M.B., Calantone, R.J., Page, T.J. and Taylor, C.R. 2000, "An Application of Multiple Group Causal Model in Assessing Cross-Cultural Measurement Equivalence," *Journal of International Marketing*, 8 (4).

Nunnally, J.C. 1978, *Psychometric Theory*. New York: McGraw-Hill.

Nunnally, J.C. and Bernstein, I.H. 1994, *Psychometric Theory*. New York: McGraw-Hill.

Onkvisit, S. and Shaw, J.J., 1989, International Marketing. Columbus, Ohio: Merrill Publishing, co.

Parameswaran, R. and Yaprak, O. 1987," A Cross-National Comparison of Consumer Research Measures," *Journal of International Business Studies*, (Spring): 35-49.

Pareek, U. and Rao, T.V. 1980, "Cross-Cultural Surveys and Interviewing," in Triandis, H.C. and Berry, J.W. *Handbook of Cross-Cultural Psychology*. Vol. 2, Boston: Allyn Bacon, 127-180.

Pérez-Rivera, M. 1994. "Dissimilar Response patterns to Likert type scale due to Cross-cultural Differences: A Comparative Study between Hispanics and Non-Hispanics Consumers" BALAS Proceeding (Business Association of Latin American Studies) 10, 375-382.

Pérez, Myra M. 1996 "Is there a Reliable Scale for Assessing Attitudes and Preferences among Hispanic and Non-Hispanic Consumers?" *Multicultural Marketing Proceedings,* 205.

Pick de Weiss, S. and Jones, D. 1981, Problemas Relacionados con la Aplicación de Cuestionarios de alternativa Fija y de Escalas de Actitudes en un País en Vía de Desarrollo," *Revista de la Asociación Latinoamericana de Psicología Social*, 1(1), 57-62.

Reynolds, T.J. and Jolly, J.P. 1980, "Measuring Personal Values: An Evaluation of Alternative Methods," *Journal of Marketing Research* 17:531-36

Riordan, C. R., and Vandenberg, R. J. 1994, "A central question in cross-cultural research: Do employees of different cultures interpret work-related measures in an equivalent manner?" *Journal of Management*, 20, 643–671.

Roster, C.A., Rogers, R.D., Albaum, G. and Klein, D. 2004. "A comparison of response characteristic from web and telephone surveys," *International Journal of Market Research, 46*(3), 359–373.

Ryan, A. M., Chan, D., Ployhart, R. E., and Slade, A. L.1999, "Employee attitude surveys in a multinational organization: Considering language and culture in assessing measurement equivalence." *Personnel Psychology*, 52, 37–58.

Schaffer, B. S., and Riordan, C. M. 2003, "A review of cross-cultural methodologies for organizational research: A best-practices approach." *Organizational Research Methods*, 6(2), 169–215.

Schwartz, S. H. 1999, "A theory of cultural values and some implications for work." *Applied Psychology: An International Review*, 48, 23–48.

Singh, J. 1995, "Measurement Issues in Cross-National Research," *Journal of International Business Studies*, 26(3) 597-619.

Shaw, M.E. and Wright, J.M. 1967, *Scales for the Measurement of Attitudes*. New York: McGraw Hill Book Co.

Shermis, M.D., & Lombard D. (1999). A comparison of survey data collected by regular mail and electronic mail questionnaires. *Journal of Business & Psychology, 14*(2), 341–354.

Stanton, J.M. (1998). An empirical assessment of data collection using the internet. *Personnel Psychology, 51*(3), 709–725.

Stanton, J.L., Chandran, R. and Hernández, S. A. 1982, "Marketing Research Problem in Latin America," *Journal of the Market Research Society*, 24(2): 124-139.

Steenkamp, J.E.M. and H. Baumgartner, 1998, "Assessing Measurement Invariance in Cross-National Consumer Research," *Journal of Consumer Research*, 25 (June), 78-90.

Strauss, M. 1969, "Phenomenal Identity and Conceptual Equivalence of Measurement in Cross-National comparative Research," Journal of Marriage and the Family, (May): 233-239.

Toyne, B. and Walters, P.G. P. 1993, *Global Marketing Management*. Boston: Allyn Bacon.

Van de Vijver, F., Jr. And Leung, K. 1997, "Methods of Data Analysis and Comparative Research," in *Handbook of Cross-Cultural Psychology*, 11, 257-300.

Van de Vijver, F., Jr. And Poortinga, Y.H. 1982, "Cross-cultural Generalization and Universality," *Journal of Cross-Cultural Psychology*, 12 (4): 387-408.

Vandenberg, R. J., and Lance, C. E. 2000, "A review and synthesis of the measurement equivalence literature: Suggestions, practices, and recommendations for organisational research." *Organizational Research Methods*, 3(1), 4–70.

Van Herk, H., Poortinga, Y. H., and Verhallen, T. M. M. 2004, "Response styles in rating scales: Evidence of method bias in data from six EU countries." *Journal of Cross-Cultural Psychology*, 35, 346–360.

Wong, N. Rindfleisch, A., Burroughs, J.E., Steenkamp, J.E.M. and Bearden, W.O. 2003, "Do Reverse-Worded Items Confound Measures in Cross-Cultural Consumer Research?" *Journal of Consumer Research* 30(1) 72-92.

Worchester, R. and Downhan, J. 1991, *Consumer Market Research Handbook.*

Wyatt, R.C. and Meyers, L.S. 1987, "Psychometric Properties of Four 5 Point Likert type Response Scales," *Educational and Psychological Measurement*, 47:27-35.

Yaprak, A. and Parameswaran, R. 1984, "Reliability Measurement in Cross-National Survey Research: An Empirical Evaluation," *International Marketing Management*, 172-193.