Journal of Extension
www.joe.org

February 2012
Volume 50 Number 1
Article Number 1FEA5

Return to Current Issue

The Forgotten Half of Program Evaluation: A Focus on the Translation of Rating Scales for Use with Hispanic Populations

Shannon J. Dogan
Associate Director of 4-H Program and Policy
University of California Division of Agriculture and Natural Resources
California 4-H Youth Development Program
Davis, California
sjdogan@ucdavis.edu

Stephanie L. Sitnick
Post Doctoral Researcher
University of Pittsburgh
Pittsburgh, Pennsylvania
sls151@pitt.edu

Lenna L. Onati
Associate Specialist in Cooperative Extension
University of California, Davis
Davis, California
lontai@ucdavis.edu

Abstract: Extension professionals often work with diverse clientele; however, most assessment tools have been developed and validated with English-speaking samples. There is little research and practical guidance on the cultural adaptation and translation of rating scales. The purpose of this article is to summarize the methodological work in this area as it relates to evaluation in Extension, specifically with Spanish-speaking, Hispanic populations of Mexican origin. General practices are reviewed and recommendations outlined. Inferences about program outcomes and impacts depend, in large part, on the rating scale; therefore, inattention to these issues could lead to misleading results and interpretations.

Introduction

Hispanics are the largest and fastest growing minority group in the nation (U. S. Census Bureau, 2004). It is estimated that by year 2020 there will be almost 60 million Hispanics, totaling approximately 18% of the

population. As Extension professionals, we continually develop and adapt programs to meet the changing needs of clientele, and there are numerous articles highlighting these efforts with Hispanic audiences (e.g., Anding, Fletcher, Van Laanen, & Supak, 2001; Israelsen, Young, & Boman, 2006; Kock, 2003; Peterson et al., 2008). An important part of this work is evaluating whether programs are successful in obtaining desired outcomes and impacts; however, most of the available assessment tools have been developed and validated with English-speaking, middle class samples, resulting in measures that may not generalize to other populations. Thus, Extension professionals often either translate existing scales or develop new questionnaires to assess program outcomes and impacts.

Research on the development and cultural adaptation of surveys for use with Hispanic audiences has generally focused on the translation and meaningfulness of the individual items and process undertaken (e.g., Brislin, 1970; Carroll, Holman, Segura-Bartholomew, Bird, & Busby, 2001; Cha, Kim, & Erlen, 2007; Farner, Cutz, Farner, Seibold, & Abuchar, 2006; Sperber, Devellis, & Boehlecke, 1994; Warrix, Nieto, & Nicolay, 2006). Few studies have focused on the cultural adaptation and translation of the rating scales or response categories (e.g., Arce-Ferrer, 2006b; Clarke, 2000; Hui & Triandis, 1989) and there are no published studies in the *Journal of Extension* on this topic.

One might ask why consideration of the rating scales is important. A rating scale is a set of categories that elicit information about the direction and strength of a participant's feelings, beliefs, opinions, and knowledge. Therefore, it is important to ensure appropriate practices are followed in the development, cultural adaptation, and translation of rating scales because the validity of the data, inferences about program outcomes and impacts, as well as cross-cultural comparisons depend, in large part, on the rating scale. Thus, while previous research has clearly outlined some of the best practices and methodological issues with regard to the translation of measures, the importance of rating scales warrant greater attention in both research and practice.

In this article, research on extreme responding and techniques for reducing this response style are reviewed and recommendations for Extension professionals outlined. The focus is on Spanish-speaking, Hispanic populations of Mexican origin. This population was chosen for two reasons. First, Mexican Americans currently comprise 66% of all Hispanics in the United States (U.S. Census Bureau, 2009). Second, there is some evidence that findings may differ among other Hispanic populations from Central America, the Caribbean, and South America (e.g., Gibbons, Zellner, & Rudek, 1999). Whether the goal is to adapt and translate an existing questionnaire for cross-cultural use, develop a new assessment tool, or interpret cross-cultural findings from one's own or others' research, this article provides a guide for thinking about the forgotten half of evaluation instruments—the rating scales. Inattention to these issues could lead to misleading results and interpretations.

Background

A body of literature exists that shows differences in English- and Spanish-speakers' response styles. A response style is the behavioral tendency of selecting particular regions on rating scales regardless of the scale's content (Cronbach, 1950). This research consistently shows that English-speakers tend to overuse the middle categories (Clarke, 2000; Hui & Triandis, 1989) and Spanish-speakers of Mexican American origin exhibit a stronger tendency toward more extreme responses (i.e., responding toward the ends of the scale) or acquiescent responding (i.e., a tendency to agree with or indicate positive connotation) (Albaum, 1997; Arce-Ferrer, 2006b; Hui & Triandis, 1989; Marín, Gamba, & Marín, 1992). For example, on a 5-point response scale (1=strongly agree, 3= neither agree nor disagree, 5=strongly disagree) Spanish-speakers are more likely to respond with a "1" or "2" than English-speakers.

Several studies have investigated factors that account for differences in response styles, such as

acculturation, education, familiarity with surveys, communication norms, and meaningfulness of the items. For example, Marín et al. (1992) found that for Hispanic participants' of Mexican origin residing in San Francisco, extreme and acquiescent responding was negatively associated with acculturation and education but not gender. Another study showed that Spanish-speakers from rural schools in the southern region of Mexico exhibited more extreme responding than respondents from urban settings (Arce-Ferrer, 2006b). The authors attributed extreme responding to familiarity with Western testing procedures and the use of rating scales.

Another body of research indicates that daily and cultural communication norms influence extreme responding. Arce-Ferrer (2006b) used a rating scale task to assess Hispanic, Spanish-speaking participants' subjective categories of judgment. Respondents were given a 7-point rating scale with end points labeled as Totally Agree (i.e., Totalmente de acuerdo) and Totally Disagree (i.e., Totalmente en desacuerdo) and asked to provide the five missing intermediate labels. Responses were rated for the degree of equivalence between the participants' subjective categories and the actual scale categories. Interestingly, as equivalence increased, extreme responding decreased, and more extreme responding occurred among participants with a limited number of subjective categories. Taken together, these results suggest that Spanish-speaking respondents may use fewer categories or semantic distinctions in their daily communication than typically targeted in scales and, therefore, may be less likely to endorse intermediate categories, leading to greater responding at the extremes.

Additionally, there is some evidence that selection of culturally meaningful intermediate categories improves validity and, in turn, reduces extreme responding among Spanish-speaking, Hispanic participants of Mexican origin (Arce-Ferrer, 2006a). For example, in our own work evaluating a program implemented with English-speaking, Anglo-American, and Spanish-speaking, Hispanic participants of Mexican origin, we found that in translating the categories "A little like me" and "Somewhat like me" there was not an appropriate Spanish translation that resulted in semantically distinct categories. As a result, we adapted the 6-point scale by dropping the category "A little like me" and retaining "Somewhat like me" as a middle category. Failure to do this might have resulted in English-speakers selecting "Somewhat like me" and Spanish-speakers selecting the more extreme but semantically equivalent "A little like me" when, in fact, there were no real differences between the groups on the item. These findings illustrate the importance of selecting culturally meaningful categories that are aligned with communication norms in the target audience.

As reviewed, there are several variables that influence extreme responding and need to be considered in the development or translation of survey instruments for Spanish-speaking, Mexican American audiences. Next, we review research on various techniques researchers have employed to reduce extreme responding.

General Practices

A body of literature has tried to understand what practices reduce the extent and effect of extreme responding. Most of this research has focused on altering the number of response categories. For example, Clarke (2000) investigated the effect of the number of response categories (three, five, seven, and nine) on extreme responding in a sample of Hispanic undergraduate students from Mexico and non-Hispanic undergraduate students from Virginia.

Extreme responding was reduced for Hispanics and non-Hispanics as the scale went from three to five categories. No additional decreases were found for seven or nine categories. These results indicate that scales with five or more categories should be used with Spanish-speaking, Hispanic participants of Mexican origin.

In a similar vein, studies have investigated how the number of response categories affects extreme

responding when participants are able to provide their own labels. For example, Hui and Triandis (1989) compared male Hispanic (of Mexican origin) and non-Hispanic American Navy recruits who completed either a 5-point or 10-point questionnaire. The labels of the scale endpoints (anchors) were the same for both formats, and the participants were free to supply labels to the other categories. The results indicated that Hispanic respondents made significantly more extreme ratings than non-Hispanic respondents but only for the 5-point scale. Additionally, the frequency distributions for the 5-point format were very different for the two groups and did not have similar shapes or overlap. Conversely, the frequency distributions were much more similar and overlapped considerably for the 10-point format.

The authors argue that more extreme responding groups have more subjective categories that represent greater intensity than is typically allowed on scales and resolve this by mapping several categories onto the same response category—leading to more extreme responding. For instance, an individual with 10 subjective categories of judgment who is given a scale with only five categories would be forced to give a 5 to questions that they might otherwise respond 8, 9, or 10 to. Thus, extreme responding is reduced when participants are allowed to give their own labels to categories or instruments include 10 categories with finer gradation at the end points.

Another technique, the two-stage scale, has emerged to mitigate the tendency of English-speaking participants to overuse the middle categories. A respondent is first asked about the direction of their response (e.g., agree, disagree, neither) and then the intensity or how strongly they feel about their response on a graphical scale (e.g., slightly, strongly). In predominately white samples of European decent, participants' use of this technique is found to lead to more extreme responding than is typically seen in other studies and greater predictability (Albaum, 1997; Albaum & Murphy, 1988). Furthermore, in a study of Spanish-speaking, Hispanic respondents of Mexican origin, extreme responding was similar whether using one- or two-stage rating formats (Arce-Ferrer, 2006b). Taken together, the results of these studies suggest that the use of a two-stage scale might minimize differences in response styles between English-speaking and Spanish-speaking respondents; however, this has not been investigated.

As mentioned earlier, the literature on cultural adaptation and translation of survey instruments has focused on achieving proper grammar and linguistic accuracy; however, as this review highlights, the rating scales themselves are an important area for future research. Understanding the issues presented is the first step to producing more culturally valid measures and valid results. Based on the literature reviewed, next we present some practical recommendations for Extension professionals interested in developing questionnaires and accompanying rating scales or adapting measures for use with Hispanic, Spanish-speaking clientele of Mexican origin.

Practical Recommendations

1. Increase the number of response categories to reduce extreme responding and allow for finer gradation (i.e., more categories) at the extremes. At least five categories are recommended (Clarke, 2000), and there is support for benefits of up to 10 categories (Hui & Triandis, 1989).

2. Provide end points or anchors, but allow respondents to supply meanings to the other categories. This allows participants to supply their own meanings to categories. This reduces extreme responding as participants are able to spread their responses across the entire range of the scale. With anchors it is still possible to test the equivalence of the measure across groups or versions. (See Recommendation 7.)

3. If intermediate category labels are desired, identify the semantic distinctions used in daily communication with your clientele and use them. Studies have shown that when rating scales are

adapted to reflect features of the respondent's functional categories of judgment this improves the validity of the scores (e.g., Arce-Ferrer, 2006a).

4. Include additional variables and measures in the study to assess and, therefore, control for factors found to influence extreme response styles, such as acculturation, education, familiarity with surveys, meaningfulness of the items, and cultural and communication norms.

5. Adapt instruments and the accompanying response scales to achieve both linguistic and cultural meaningfulness. This requires more than simply translating the questions and response categories into another language.

6. Pilot measures with the target clientele to assess whether they are function as expected. We recommend collecting data from the target group using the adapted instrument in order to examine the psychometric properties (i.e., reliability, construct validity, content validity, predictive validity, etc.) (McDonald, 1999). It may be desirable to administer the original instrument too. (See Recommendation 7.) Qualitative assessments such as focus groups can also help uncover translation issues and categories that are functionally equivalent and not semantically distinct.

7. Appropriate statistical techniques need to be applied to establish measurement equivalence (invariance) of the adapted version either across groups or in relation to the original instrument—depending on the goal (Byrne, Shavelson, & Muthén, 1989; Cheung & Rensvold, 2000; Widaman & Reise, 1997). Equivalence of measures across groups is a requirement if one wants to conduct cross-cultural comparisons of program effects. Equivalence of the original and adapted version of the instrument is necessary to assure the adapted measure functions similarly to the (validated) original measure and to make comparisons to studies that used the original measure.

8. Be aware of factors that influence extreme responding and recognize them during the development and adaptation of surveys, as well as during data analysis and interpretation.

Conclusion

U.S. Extension professionals who have experience working with diverse clientele know that we cannot simply take a program that works with one population and assume it will work in the same way with another population. Likewise, the studies reviewed in this article demonstrate that we cannot simply translate a response scale that is shown to be reliable and valid in one cultural context and assume it will function in the same manner in a different cultural context.

Care must be taken in the development of assessment tools, and this includes more than just attention to the items in the scale. Failure to consider response categories and factors that underlie extreme responding may result in program outcomes, impacts, and cross-cultural comparisons that are invalid. More research is needed in this area; however, by being aware of and using these recommendations, Extension professionals can improve the development and adaptation of survey instruments for use with Spanish-speaking, Hispanic populations of Mexican origin and their evaluation of program outcomes and impacts.

References

Albaum, G. (1997). The Likert scale revisited: An alternate version. *Journal of the Market Research Society*, 39(2), 331-348.

Albaum, G., & Murphy, B. D. *(1988). Extreme response on a Likert scale*. Psychological Reports, 63(2),

501-502.

Anding, J., Fletcher, R. D., Van Laanen, P., & Supak, C. (2001). The Food Stamp Nutrition Education Program's (FSNEP) impact on selected food and nutrition behaviors among Texans. *Journal of Extension* [On-line], 43(1) Article 6RIB4. Available at: http://www.joe.org/joe/2001december/rb4.php

Arce-Ferrer, A. J. (2006a). Investigating with IRT and MDS approaches translation and adaptation of rating scales for Spanish-speaking populations. *International Journal of Testing*, 6(3), 269-285.

Arce-Ferrer, A. J. (2006b). An investigation into the factors influencing extreme-response style: Improving meaning of translated and culturally adapted rating scales. *Educational and Psychological Measurement*, 66(3), 374-392.

Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of Cross Cultural Psychology*, 1(3), 185-216.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, 105, 456-466.

Carroll, J. S., Holman, T. B., Segura-Bartholomew, G., Bird, M. H., & Busby, D. M. (2001). Translation and validation of the Spanish version of the RELATE questionnaire using a modified serial approach for cross-cultural translation. *Family Process*, 40(2), 211-231.

Cha, E.-S., Kim, K. H., & Erlen, J. A. (2007). Translation of scales in cross-cultural research: Issues and techniques. Journal of Advanced Nursing, 58(4), 386-395.

Cheung, G. W., & Rensvold, R. B. (2000). Assessing extreme and acquiescence response sets in cross-cultural research using structural equations modeling. *Journal of Cross-Cultural Psychology*, 31(2), 187-212.

Clarke, I., III. (2000). Extreme response style in cross-cultural research: An empirical investigation. *Journal of Social Behavior & Personality*, 15(1), 137-152.

Cronbach, L. J. (1950). Further evidence on response sets and test design. *Educational and Psychological Measuremen*t, 10, 3-31.

Farner, S., Cutz, G., Farner, B., Seibold, S., & Abuchar, V. (2006). Running successful extension camps for Hispanic children: From program planning to program delivery for a 1-week day camp. J*ournal of Extension* [On-line], 44(4) Article 4FEA4. Available at: http://www.joe.org/joe/2006august/a4.php

Hui, C. H., & Triandis, H. C. (1989). Effects of culture and response format on extreme response style. *Journal of Cross-Cultural Psychology*, 20(3), 296-309.

Gibbons, J. L., Zellner, J. A., & Rudek, D. J. (1999). Effects of language and meaningfulness on the use of extreme response style by Spanish-English bilinguals. *Cross-Cultural Research: The Journal of Comparative Social Science*, 33(4), 369-381.

Israelsen, C. E., Young, A. J., & Boman, R. L. (2006). Milking and calf care schools for Hispanics in Cache Country. *Journal of Extension* [On-line], 44(4) Article 4IAW2. Available at: http://www.joe.org/joe/2006august/iw2.php

Kock, J. A. (2003). Children's literacy: Children's books for healthy families/libros de ninos para familias

saludables. *Journal of Extension* [On-line], 41(2) Article 2FEA7. Available at: http://www.joe.org/joe/2003april/a7.php

Marín, G., Gamba, R. J., & Marín, B. V. (1992). Extreme response style and acquiescence among Hispanics: The role of acculturation and education. *Journal of Cross-Cultural Psychology*, 23(4), 498-509.

McDonald, R. P. (1999). Test theory: A unified treatment. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Peterson, S. S., Gerstein, D. E., Mugford, K., Wiley, R., Davis, J., Nicholson, L., et al. (2008). Eat Smart. Play Hard San Luis Obispo: A nutrition and fitness pilot program for young children and their adult buddies. *Journal of Extension* [On-line], 46(1) Article 1IAW1. Available at: http://www.joe.org/joe/2008february/iw1.php

Sperber, A. D., Devellis, R. F., & Boehlecke, B. (1994). Cross-cultural translation: Methodology and validation. *Journal of Cross-Cultural Psychology*, 25(4), 501-524.

U. S. Census Bureau (2004). U.S. interim projections by age, sex race, and Hispanic origin. Retrieved from: http://www.census.gov/population/www/projections/usinterimproj/natprojtab01a.pdf.

U. S. Census Bureau (2009). American Community Survey. Retrieved from: http://factfinder.census.gov/servlet/DTTable?_bm=y&-geo_id=01000US&-ds_name=ACS_2009_1YR_G00_&-_lang=en&-redoLog=true&-mt_name=ACS_2009_1YR_G2000_B03001&-format=&-CONTEXT=dt

Warrix, M. B., Nieto, R. D., & Nicolay, M. (2006). Developing culturally appropriate evaluation instruments for Hispanics with diabetes. *Journal of Extension* [On-line], 44(6) Article 6TOT1. Available at: http://www.joe.org/joe/2006december/tt1.php

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In J. K. Bryant, M. Windle & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281-324). Washington, DC: American Psychological Association.
