Quantitative Evaluation of Response Scale Translation Through a Randomized Experiment of Interview Language With Bilingual English- and Spanish-Speaking Latino Respondents

Sunghee Lee, Mengyao Hu, Mingnan Liu, and Jennifer Kelley

Introduction

Survey data collection using multiple languages has increased dramatically with a greater interest in research concerning populations that speak different languages. Questionnaire translation, once viewed as only integral for international surveys (e.g., Ervin & Bower, 1952), is now needed even for surveys within a single country. In the United States, for example, it has become a standard practice to conduct surveys in both English and Spanish languages for scientific population-based data collection. Spanish has become a standard interview language in the United States for two reasons. First, the number of Latinos living in the United States has increased sharply. Persons reporting Latino origin grew from 35.3 million to 50.5 million between 2000 and 2010, corresponding to 13 and 16 percent of the total US population, respectively (Ennis, Ríos-Vargas, & Albert, 2011). What sets Latinos apart from non-Latinos is their language use. According to the 2010 American Community Survey, close to 8 out of 10 Latinos aged 5 years or older spoke Spanish at home. Among those who spoke Spanish at home, nearly half reported speaking English less than "very well," which the US Census Bureau uses as a working definition of "linguistically isolated" (Ryan, 2013; Siegel, Martin, Bruno, Martin, & Siegel, 2001; see Chapter 3 for background information on the term "linguistically isolated," now referred to as "limited English speaking"). Second, English proficiency is associated with various educational, economic, health, and social behaviors (Institute of Medicine, 2003; Yu, Nyman, Kogan, Huang, & Schwalberg, 2004). Hence, interviewing only in English incurs unexpected incorrect representation of the US population (Korey & Lascher, 2006; Lee, Nguyen, Jawad, & Kurata, 2008).

While conducting interviews in multiple languages improves the scope of the population covered in a given survey, it also introduces challenges to the measurement properties that are not present in monolingual surveys (Smith, 2009). In a multilingual survey, the differences in responses across languages may reflect not only true differences in the concept that a question seeks to measure but also measurement artifacts due to translation. This chapter introduces a way to evaluate the translation of response scales using an experiment implemented in a questionnaire targeting bilingual English- and Spanish-speaking Latino respondents in the United States.

Translation and Measurement Equivalence

Translation is a necessary and crucial step in multilingual surveys. In most translation practices, a questionnaire is prepared in one language (source language) and then translated into other languages (target languages) (Harkness, 2003). Given that languages are not isomorphic, translation is more than a mechanical process that finds semantically and lexically close texts. It often involves careful adaptation for use in the cultures associated with the target languages. The rationale behind this practice is to retain measurement properties equivalent across languages. Measurement equivalence in multilingual surveys can be described in many ways. For example, Johnson (1998, Table 1) lists 52 types of equivalence ranging from vocabulary equivalence to theoretical equivalence. In this chapter, we use *functional equivalence* to describe measurement equivalence. Per Scheuch (1968), functional equivalence extends beyond comparability in the meaning and implies equivalence for the purpose of analysis. When a question is not functionally equivalent between the source and target languages, the measured construct or concept may not be comparable.

Translation may hamper measurement equivalence in multilingual surveys by affecting respondents' cognitive processes when answering questions. More specifically, translation may affect how respondents interpret the questions, what information they retrieve from their memories, how they use the retrieved information for rendering the appropriate judgment, and finally how they map their judgment onto the response scales (Yan & Hu, 2018).

Translation of Response Scales

Because response scales are closely tied to respondents' cognitive processes, translation of *response scales* is of critical importance (Mohler, Smith, & Harkness, 1998). Given that respondents may perceive the meaning or magnitude of a specific response category in a given response scale specific to each language, translation may affect how respondents interpret and map their answers onto the scale. Because respondents may use the response scales presented with questions to help interpret the meaning of the questions, response scale translation may also affect how respondents understand the questions. Overall, lack of measurement equivalence introduced by response scale translation is likely to distort the response distribution, making analysis noncomparable (Keller et al., 1998).

For a target language, there is no consensus on how to effectively translate response scales. In fact, the extant literature includes frequent observations in which, for a given response scale in the source language, various versions exist in the same target language. The difficulty of translating response scales has been explicitly reported for the Likert agreement scale in Japanese, German, and Swahili. For example, Shishido, Iwai, & Yasuda (2009) reported that "agree" and "disagree" have been translated as *sansei* ("agree") and *hantai* ("disagree") and as *sou omou* ("I think so") and *sou omowanai* ("I don't think so") in Japanese surveys and that Japanese respondents expressed their opinions more clearly on *sou omou* ("I think so") and *sou omowanai* ("I don't think so") than on the other versions. German does not offer a formally matched expression of "disagree"; Hebrew and Swahili do not have a well-matched expression of "neither agree nor disagree" (Harkness, Pennell, & Schoua-Glusberg, 2004; Harkness, Villar, & Edwards, 2010; Yan & Hu, 2018). Similar difficulties are reported for the "excellent-very good-good-fair-poor" response scale, where response categories in a source language are translated differently depending on the target language.

Yan and Hu (2018) examined translations of the "excellent" to "poor" scale in several national surveys. They found that the category "fair" was translated as 一般 ("average") in Chinese, *mittelmäßig* ("middle" or "mediocre") in German, and *ganska dålig* ("somewhat poor") in Swedish, resulting in incomparable results across cultures. Although difficulties of translating response categories are not widely reported for Spanish, some researchers discuss response categories as a source of noncomparability in reports between Latino and non-Latino respondents in the United States (Bzostek, Goldman, & Pebley, 2007; Kandula, Lauderdale, & Baker, 2007; Viruell-Fuentes, Morenoff, Williams, & House, 2011) and sensitivity of the Likert scale presentation in Spanish (Arce-Ferrer, 2006). Response scale translation may also change the structure of the scales that respondents perceive implicitly (e.g., changing unipolar into bipolar scales and changing balanced scales into unbalanced scales). For example, for the self-rated health question using an "excellent-very good-good-fair-poor" scale, "poor" has been translated into a word meaning "not good" in some surveys and "bad" in other surveys using the same target language (Behr, Dept, & Krajčeva, 2018). As respondents assign meanings to numeric values (Schwarz, Knauper, Hippler, Noelle-Neumann, & Clark, 1991), if we match the translated response categories to numbers, "not good" could be understood as zero on a unipolar scale of goodness, while "bad" could be understood as a negative value on a bipolar scale of bad to good (Yan & Hu, 2018). This structural change may bias the survey estimates because "poor" actually means worse health when translated into a word meaning "bad" rather than "not good."

Translation Evaluation

There are various approaches for evaluating questionnaire translation as discussed in the Cross-Cultural Survey Guidelines published by the University of Michigan. Qualitative approaches, such as experts' review, feedback from translators, cognitive interviews, and behavioral coding (e.g., Dept, Ferrari, & Wäyrynen, 2010; Gordoni & Schmidt, 2010; Hunt & Bhopal, 2004; Willis et al., 2010), are commonly used. Qualitative approaches are the necessary first step to ensuring translation quality, and their dominance reflects practical constraints on resources in survey research (Tourangeau, 2004). Translation evaluation can also take a quantitative approach, which may provide a higher level of generalizability and reproducibility (Harkness et al., 2004). However, quantitative research on translation is rather sparse.

Quantitative approaches for assessing translation can be classified into two categories: (1) experiments designed to collect assessment data and (2) statistical models with existing data. Most quantitative studies use the latter (e.g., Davidov & De Beuckelaer, 2010; Saris, 2003; also see Braun & Johnson, 2010; Van de Vijver, 2003; and Van de Vijver & Leung, 1997 for an overview of the modeling approaches). Data for statistical models may but typically do not involve randomized experiments on translation. While conceivable, experiments with bilingual respondents who are fluent in both source and target languages have been rarely used for translation evaluation (Smith, 2004). When these bilingual respondents are randomly assigned to either language for a survey interview, they are comparable except for the interview language. Hence, equivalence between source and target languages can be tested directly by comparing estimates between languages. Moreover, if there are multiple versions of translation of a particular response scale in a target language, they can also be assessed to compare their levels of equivalence with the source language.

Goal of This Research

To address the need to evaluate response scale translation quantitatively, this chapter uses data from an experiment on interview language conducted in a population-based survey that targeted racial and ethnic minorities in the United States. The interview language experiment was implemented for bilingual Latinos who reported speaking English and Spanish about the same amount of time, providing unique data that allow us to examine measurement equivalence in translated questionnaires quantitatively.

We focused on the translation of quantifier-based ordinal response scales. As noted earlier, translation of these response scales is difficult because they combine both negation and quantification, and the available lexical and structural options for the scales differ across languages (Harkness et al., 2004). Moreover, when translated, the vagueness of quantifiers may elicit nonequivalent measurement structures.

Data and Method

Data Source

We used data from the National Latino and Asian American Study (NLAAS) fielded between May 2002 and November 2003. NLAAS was conducted specifically to overcome the lack of population-based data for Latino and Asian Americans in the United States. Targeting adults aged 18 years old or older in those racial and ethnic groups, the study used a stratified area-probability sampling. To account for high linguistic isolation rates of the target population, NLAAS interviews were conducted in Spanish, Chinese, Vietnamese, and Tagalog in addition to English by fully bilingual interviewers. The questionnaire was first developed in English and translated into other languages. The sample comprised 2,554 Latino and 2,095 Asian American adults. Pennell et al. (2004) and Takeuchi, Gong, and Gee (2012) offered detailed accounts of NLAAS and Alegria et al. (2004) of cultural adaptation and translation processes in NLAAS.

At the beginning of the interview, Latino respondents were asked about their English and Spanish usage. Among them, 827 reported speaking only Spanish, 521 mostly Spanish, 332 Spanish and English about the same amount of time, 627 mostly English, and 227 only English. NLAAS regarded those 332 who reported speaking English and Spanish about the same amount of time as bilingual and randomly assigned them to either Spanish or English for interviews. As a result, 182 bilingual Latino respondents completed interviews in English and 150 in Spanish. This study used data from this interview language experiment. Note that this experiment was implemented only for bilingual Latino respondents.

There were two types of translation for response scales in NLAAS. The first involved translating a scale in English into one version in Spanish. The second type translated a scale in English into two versions in Spanish. (Note that it is unclear from the NLAAS documents whether two Spanish versions for one English scale were designed intentionally.) We labeled the former as "one-on-one translation" and the latter as "one-on-two translation." Most response scales in NLAAS followed one-on-one translation. We chose four response scales in this study for two reasons. First, they are widely used in questionnaires in general. Second, each of the chosen scales was used for multiple questions on the same topic. Having multiple items reduces the chance of misinterpreting an attribute of a single item as evidence for translation equivalence and provides more analysis options.

Under one-on-one translation, we examined two response scales: (1) a 4-point excellent-to-poor scale that translated "excellent-good-fair-poor" into *excelente-bien-regular-pobre* and was used for a set of six language proficiency questions and (2) a 4-point Likert agreement scale that translated "strongly agree-somewhat disagree-strongly disagree" into *mayormente de acuerdo-algo de acuerdo-algo en desacuerdo-mayormente en desacuerdo* and was used for 10 family cohesion questions.

Two response scales fell under the one-on-two translation: (1) a 4-point frequency scale and (2) a 4-point quantity scale. The frequency scale of "often-sometimes-rarely-never" was translated into either *muchas veces-alguna veces-casi nunca-nunca* or *muchas veces-alguna veces-pocas veces-nunca*, using different Spanish words (*casi nunca* or *pocas veces*) for "rarely." The version with *casi nunca* was used for four questions about demands by social networks, while the version with *pocas veces* was used for four immigration and discrimination questions. The English version of the quantity scale was "a lot-some-a little-not at all" and was translated into either *mucho-algo-poco-nada* or *mucho-regular-poco-nada*. "Some" was translated into either *algo* or *regular*. The version with *algo* was used for seven questions on the effects of a terrorist attack, and *regular*was used for four questions about reliance on social networks. With the one-on-two translation, we can examine not only translation equivalence but also comparability in equivalence across translation versions. See Appendix 4-1 for the wording of the questions used in the study. Alegria et al. (2004) documented the backgrounds on how these questions were developed for NLAAS.

Analysis Plan

We analyzed each response scale separately. We first compared response distributions by interview language for each scale and by different Spanish translation versions for the one-on-two translation scales. Similar response distributions between English and Spanish indicate translation equivalence in the first comparison. With one-on-two translation scales, similarities in response distributions between two versions of the Spanish response scales imply that the two translated versions are comparable regardless of their individual equivalence to the English scale. For this, the relative difference in each response category was calculated by dividing the difference in estimates between Spanish and English interviews by the estimates based on English interviews and compared between the two Spanish versions. The Spanish version with smaller relative differences was considered to be more equivalent to the English version. We used a relative difference rather than an absolute difference because the latter does not provide as much information when the response distributions are uneven across response categories (e.g., skewness toward one end or concentration around one category) and illustrates the impact less clearly.

Because each scale was used for multiple topically related questions, we also computed Cronbach's α on each response scale for each language and compared it between interview languages through χ^2 tests, as illustrated in Feldt, Woodruff, and Salih (1987). If translation retained the equivalence, Cronbach's α should not be different between English and Spanish. We also conducted analysis of variance (ANOVA), suggested by Van de Vijver and Leung (1997) as an extension of Cleary and Hilton (1968). This method detects item bias caused by translation. For the ANOVA analysis, we first created a score summary variable for each scale in three steps: summed responses of all topically related items into a total score within a respondent, computed the quartile of the summary score, and assigned each respondent to a quartile. Hence, the score summary variable has four levels. We then modeled responses of each item on two main effects—the interview language and the score summary variable—as well as their interaction. In these models, the score summary variable was not of interest because individual item scores were part of the total score. Instead, the effect of the language was of interest because interview language should not play a role in explaining the variance of individual item scores due to its random assignment. If interview language was significant in the estimated model, it would indicate lack of translation equivalence. This ANOVA approach allowed us to test whether interview language contributed to the variance of the individual item scores, while controlling for the person's standing in the total score. Note that Cronbach's α and the ANOVA approach described here were feasible because each response scale had multiple items on the same topic.

Because sample sizes were relatively small, the focus of the study was not necessarily to detect statistical significance. Rather, it was to demonstrate how such experimental data can be used for evaluating a translation quantitatively. We attempted to understand potential changes in measurement due to translation with commonly used response scales and, when more than one translation version was used, to propose a better version. Because of the experimental nature of the data, the results presented here did not consider population-level weight adjustments.

We note that the randomization of interview language should have produced two groups of respondents with similar characteristics. In comparing sociodemographic characteristics, specifically, age (18–30 years old, 31–50 years old, 51 years old or older), gender (male, female), education (less than high school, high school, some college, college or more), nativity (US born, foreign born), and Latino subgroups (Mexican, Puerto Rican, others), we found most were comparable between the English and Spanish interview language groups. However, the proportion of the age category 18–30 years was not even; there was a larger proportion in the English interview groups compared with the Spanish interview groups (44.0 percent vs. 34.0 percent, $p = .035$, respectively). This discrepancy led us to assume an uneven breakoff pattern by younger respondents interviewed in Spanish. The smaller sample size of the Spanish interviews compared with the English interviews (150 vs. 182) may be indirect evidence. Because there is no information about the breakoffs in the NLAAS data or documents, this assumption was not verified. Instead, to maintain the comparability, we adjusted for any potential differences between language groups with respect to the previously listed characteristics in all analyses by standardizing their marginal distributions using the English group as a benchmark. All analyses were conducted in SAS, except for the comparison of Cronbach's α, which used an R package "cocron" (Diedenhofen & Musch, 2016).

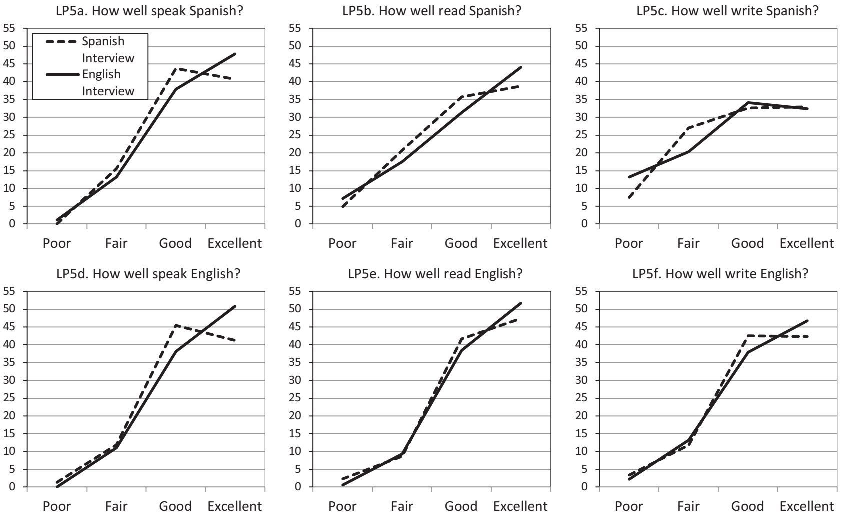
Results

One-on-One Translation

Excellent-to-Poor Scale

How bilingual Latino respondents rated their own speaking, reading, and writing aspects of Spanish and English language proficiency is presented by interview language in Figure 4-1. For all measures except the Spanish writing aspect, respondents interviewed in English chose "excellent" at a consistently higher rate than those interviewed in Spanish. This choice made those interviewed in English appear more proficient in both English and Spanish, even though, in reality, these respondents were comparable in their language use. Although we do not discuss this response scale in this chapter, it is notable that the same pattern emerged for questions on physical and mental health, which used a 5-point excellent-to-poor scale ("excellent-very

Figure 4-1. Distribution of Spanish and English proficiency on speaking, reading, and writing, by interview language



good-good-fair-poor" translated into *excelente-muy-bien-bien-regular-pobre*): bilingual respondents interviewed in English chose "excellent" and "very good" categories at a higher rate than those interviewed in Spanish, making English-language respondents look as though they were healthier than Spanish-language respondents (results not shown).

Given that speaking, reading, and writing aspects all measure the concept of language proficiency, they should be related for a given language. To test this idea, we compared Cronbach's α by interview language. Cronbach's α for Spanish proficiency measured higher among those interviewed in Spanish at .913, compared with .885 among those interviewed in English, but the difference was not statistically significant ($\chi^2 = 1.56$ [*df* = 1]; $p = .212$). For English proficiency measures, Cronbach's α was comparable at .930 and .938 for the Spanish and English interviews, respectively. In the ANOVA models, interview language was significant in explaining English speaking scores as a main effect as well as through an interaction with the score summary. The English reading score was higher for bilingual Latino respondents who were interviewed in English rather than in Spanish. (See Appendix 4-2 for detailed results of all ANOVA models.)

Agreement Scale

On the 4-point agreement scale used for 10 family cohesion questions, the "strongly agree" category was chosen most frequently for both interview languages. However, this tendency was more pronounced for Spanish than English interviews, as shown by comparing proportions of "strongly agree" between languages in Table 4-1. Even with the small sample size, language of interview was significant at $p < .05$ for questions such as "Things work well for us as a family (FC3)" and "We really do trust and confide in each other (FC4)," for which Spanish interviewees used "strongly agree" by 14.8 and 11.5 percentage points higher than English interviewees, respectively, and at $p < .1$ for "We share similar values and beliefs as a family (FC2)" and "Family togetherness is very important (FC10)," with 9.2 and 8.2 percentage point differences, respectively.

Cronbach's α across family cohesion questions was not significantly different between interview languages (.931 for English and .929 for Spanish). Language in ANOVA introduced earlier showed a significant effect on one item (FC3) through an interaction ($p = .016$). Among those in the third and fourth quartiles of the total score, those interviewed in Spanish showed a significantly higher score on this item than those interviewed in English.

Table 4-1. Proportion of "strongly agree" for family cohesion questions, by interview language

Question: Now I'd like to know how strongly you agree or disagree with the following statements about your family.	**Interview Language**		**Difference: Spanish–English**	***p* value**
	Spanish % (*SE*)	**English % (*SE*)**		
	***n* = 149**	***n* = 182**		
FC1. Family members respect one another.	70.0 (4.0)	63.7 (3.6)	6.2	.244
FC2. We share similar values and beliefs as a family.	69.1 (4.0)	59.9 (3.6)	9.2	.085
FC3. Things work well for us as a family.	69.7 (4.0)	54.9 (3.7)	14.8	.007
FC4. We really do trust and confide in each other.	71.9 (3.9)	60.4 (3.6)	11.5	.031
FC5. Family members feel loyal to the family.	74.0 (3.9)	66.5 (3.5)	7.5	.149
FC6. We are proud of our family.	82.5 (3.2)	75.8 (3.2)	6.7	.139
FC7. We can express our feelings with our family.	66.6 (4.1)	61.5 (3.6)	5.0	.357
FC8. Family members like to spend free time with each other.	59.5 (4.3)	52.2 (3.7)	7.3	.194
FC9. Family members feel very close to each other.	67.1 (4.1)	65.9 (3.5)	1.2	.830
FC10. Family togetherness is very important.	81.9 (3.3)	73.6 (3.3)	8.2	.078

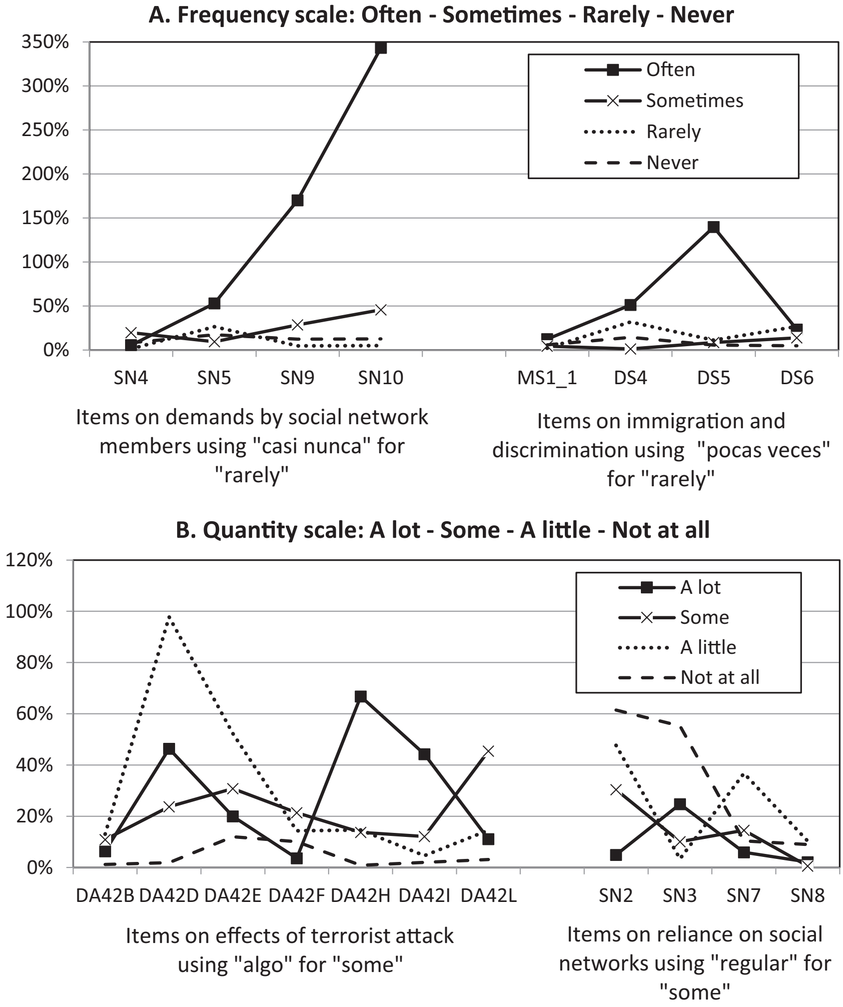
One-on-Two Translation

Frequency Scale

For the frequency scale of "often-sometimes-rarely-never" where "rarely" was translated into two Spanish versions, *casi nunca* and *pocas veces*, we examined the relative difference for each response category between the English version and each Spanish version and compared the relative differences between the two Spanish versions in Figure 4-2A. The differences were particularly large for the "often" and "sometimes" categories with the Spanish scale using *casi nunca* rather than *pocas veces*. The average of the question-level relative difference was 41.0 percent with *casi nunca* compared with 22.3 percent with *pocas veces*.

While Cronbach's α was not comparable between languages when using *casi nunca* ($\alpha = .688$ vs. $\alpha = .545$ for Spanish and English, respectively; $\chi^2 = 3.41$ [$df = 1$]; $p = .064$), it was comparable with *pocas veces* ($\alpha = .729$ vs. $\alpha = .717$ for Spanish and English, respectively). From ANOVA, the interview language and its interactions with the score summary variable showed a significant effect on three of the four items using *casi nunca* (SN5, SN9, and

Figure 4-2. Percentage relative difference for items with frequency and quantity scales between Spanish and English interviews, by Spanish translation version



A. Frequency scale: Often - Sometimes - Rarely - Never

B. Quantity scale: A lot - Some - A little - Not at all

SN10), suggesting item bias due to translation. However, none of the items using the scale with *pocas veces* was subject to a significant language effect.

Quantity Scale

Eleven questions used the "a lot-some-a little-not at all" quantity scale, for which "some" was translated into either *algo* or *regular*. The relative difference reported in Figure 4-2B was consistently larger for the Spanish response scale using *algo* than the scale using *regular*. The overall mean of the relative difference was 31.1 percent for the scale with *algo* and 20.5 percent for the scale with *regular*. The difference in Cronbach's α between English and Spanish interview languages was significant for questions using *algo* ($\alpha = .649$ vs. $\alpha = .758$ for Spanish and English, respectively; $\chi^2 = 4.25$ [$df = 1$]; $p = .039$) but not for *regular* ($\alpha = .678$ vs. $\alpha = .702$ for Spanish and English, respectively). However, based on ANOVA, language showed a significant effect on one item using *algo* (DA42b) as a main effect and one item using *regular* only through its interaction with the score summary variable (SN3).

Discussion

Our analysis illustrates an assessment of measurement equivalence between English and Spanish questionnaires through an experiment that randomized interview language with bilingual English- and Spanish-speaking Latino Americans. Overall, the results show a language effect. On the "excellent-good-fair-poor" scale used for language proficiency questions, bilingual Latinos chose positive responses more frequently when interviewed in English than in Spanish. When interviewed in English, bilingual Latinos' language proficiency in both English and Spanish appeared higher. Clearly, the translated Spanish response scales did not align with the English scale on the continuum of true language proficiency. It could be that *excelente* in Spanish conveys a more desirable state than "excellent" in English.

With the agreement scale used for family cohesion questions, bilingual Latinos reported "strongly agree" at a consistently higher rate when interviewed in Spanish than in English. This trend may be related to extreme response style (ERS). It is hypothesized in the literature that Latinos are more engaged in ERS than non-Latino whites (Hui & Triandis, 1989; Marín, Gamba, & Marín, 1992; Weech-Maldonado, Elliott, Oluwole, Schiller, & Hays, 2008). While our study included only Latinos, it is imaginable that the ERS tendency of Latinos is partially due to the priming effect of the interview language. That is, when interviewed in Spanish as opposed to in English, bilingual Latinos are more likely to exhibit ERS because the Spanish language itself activates Latino-specific cultural norms promoting ERS. Further, the nature of the topic, family cohesion, is more culturally salient to Latinos than non-Latino whites because of *familismo*, one of the important Latino cultural values (Marín & Marín, 1991; Toro-Morn, 2012; Zea, Quezada, & Belgrave, 1993). Therefore, Latino cultural norms associated with the Spanish language may have influenced how bilingual Latinos responded to questions about family cohesion when these questions were asked in Spanish.

For the "often-sometimes-rarely-never" frequency scale or the "a lot-some-a little-not at all" quantity scale, this study offers quantitative evidence for better translations in Spanish. Between *casi nunca* and *pocas veces* in place of the English category "rarely," the scale with *pocas veces* produced more similar results to English than the scale with *casi nunca*. When choosing a Spanish quantifier for "some" on the "a lot-some-a little-not at all" scale, *regular* appeared somewhat more advantageous for measurement comparability than *algo*.

Of course, for the reasons behind the lack of translation equivalence shown in this chapter, one may argue that bilingual respondents bring in different cultural norms associated with the language they are interviewed in because language primes respondents' cognition (Bond, 1983; Marian & Kaushanskaya, 2004; Ross, Xun, & Wilson, 2002; Trafimow, Silverman, Fan, & Fun Law, 1997; Triandis, Davis, Vassiliou, & Nassiakou, 1965). Research has shown that bilingual people process information differently than monolingual people (Holmes, 2008), which makes it reasonable to conclude that the effect shown in this chapter may be caused by cultural differences combined with linguistic differences. In fact, the purpose of this study was not to distinguish these two. Instead, the interview language effect can be seen as a result of translation, which may activate respondents' cultural norms when they answer survey questions.

Translation is an inherent task for cross-cultural and cross-national research and is a topic that has received much attention from cross-cultural survey researchers. Unfortunately, despite the importance and broad impact, there are many inconsistent translations with no clear guidelines. Still, translation is mostly assessed through qualitative approaches. Smith (2004) recommended quantitatively evaluating the qualitative translation to ensure measurement comparability, which, in turn, lowers the chances of producing misleading results in cross-cultural studies. Similarly, Scheuch (1968) argued that literal equivalence achieved through qualitative translation procedures may not guarantee functional equivalence. This study demonstrated how experimental data with bilingual speakers provide quantifiable and objective evidence, which can enhance translation procedures.

This study has several important implications. First, it shows the importance of response scale translation and its unintended negative effects on measurement equivalence. Direct comparisons of estimates between interview languages may lead to biased results. Second, it shows difficulties with response scale translation. Inconsistent translations (e.g., *algo* or *regular* for "some") can lead to different response distributions. Third, it suggests better translation of some response scales. For instance, "some" on a frequency scale may be better translated using *regular* rather than *algo* in Spanish questionnaires when targeting US Latinos.

Other developments are underway to quantitatively assess translation and to make appropriate adjustments. Approaches such as anchoring vignettes (e.g., Hopkins & King, 2010; Hu, Lee, & Xu, 2018; Van Soest, Delaney, Harmon, Kapteyn, & Smith, 2011), item response theory (e.g., Azocar, Areán, Miranda, & Muñoz, 2001; Ellis, Minsel, & Becker, 1989), and unfolding models (e.g., Javaras & Ripley, 2007) are great examples. If using these approaches, evaluations need to be preplanned because they require specific types of data.

References

Alegria, M., Takeuchi, D., Canino, G., Duan, N., Shrout, P., Meng, X.-L., … Gong, F. (2004). Considering context, place and culture: The National Latino and Asian American Study. *International Journal of Methods in Psychiatric Research*, *13*(4), 208–220. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15719529

Arce-Ferrer, A. J. (2006). An investigation into the factors influencing extreme-response style. *Educational and Psychological Measurement*, *66*(3), 374–392. https://doi.org/10.1177/0013164405278575

Azocar, F., Areán, P., Miranda, J., & Muñoz, R. F. (2001). Differential item functioning in a Spanish translation of the Beck Depression Inventory. *Journal of Clinical Psychology*, *57*(3), 355–365.

Behr, D., Dept, S., & Krajčeva, E. (2018). Documenting the survey translation and monitoring process. In T. P. Johnson, B.-E. Pennell, I. Stoop, & B. Dorer (Eds.), *Advances in comparative survey methods: Multinational, multiregional, and multicultural contexts (3MC)* (p. 457–476). Hoboken, NJ: John Wiley & Sons.

Bond, M. H. (1983). How language variation affects inter-cultural differentiation of values by Hong Kong bilinguals. *Journal of Language and Social Psychology*, *2*(1), 57–66. https://doi.org/10.1177/0261927X8300200104

Braun, M., & Johnson, T. P. (2010). An illustrative review of techniques for detecting inequivalences. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, P. Ph. Mohler, B.-E. Pennell, & T. W. Smith (Eds.), *Survey methods in multinational, multiregional, and multicultural contexts* (pp. 373–393). Hoboken, NJ: John Wiley & Sons.

Bzostek, S., Goldman, N., & Pebley, A. (2007). Why do Hispanics in the USA report poor health? *Social Science & Medicine*, *65*(5), 990–1003. https://doi.org/10.1016/J.SOCSCIMED.2007.04.028

Cleary, T. A., & Hilton, T. L. (1968). An investigation of item bias. *Educational and Psychological Measurement*, *28*(1), 61–75. https://doi.org/10.1177/001316446802800106

Davidov, E., & De Beuckelaer, A. (2010). How harmful are survey translations? A test with Schwartz's human values instrument. *International Journal of Public Opinion Research*, *22*(4), 485–510. https://doi.org/10.1093/ijpor/edq030

Dept, S., Ferrari, A., & Wäyrynen, L. (2010). Developments in translation verification procedures in three multilingual assessments: A plea for an integrated translation and adaptation monitoring tool. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, P. Ph. Mohler, B.-E. Pennell, & T. W. Smith (Eds.), *Survey methods in multinational, multiregional, and multicultural contexts* (pp. 157–173). Hoboken, NJ: John Wiley & Sons.

Diedenhofen, B., & Musch, J. (2016). cocron: A web interface and R package for the statistical comparison of Cronbach's alpha coefficients. *International Journal of Internet Science*, *11*(1), 51–60.

Ellis, B. B., Minsel, B., & Becker, P. (1989). Evaluation of attitude survey translations an investigation using item response theory. *International Journal of Psychology*, *24*(6), 665–684. https://doi.org/10.1080/00207598908247838

Ennis, S. R., Ríos-Vargas, M., & Albert, N. G. (2011). *The Hispanic population: 2010*. Retrieved from http://www.census.gov/prod/cen2010/briefs/c2010br-04.pdf

Ervin, S., & Bower, R. T. (1952). Translation problems in international surveys. *Public Opinion Quarterly*, *16*(4), 595. https://doi.org/10.1086/266421

Feldt, L. S., Woodruff, D. J., & Salih, F. A. (1987). Statistical inference for coefficient alpha. *Applied Psychological Measurement*, *11*(1), 93–103. https://doi.org/10.1177/014662168701100107

Gordoni, G., & Schmidt, P. (2010). The decision to participate in social surveys: The case of the Arab minority in Israel—An application of the theory of reasoned action. *International Journal of Public Opinion Research*, *22*(3), 364–391. https://doi.org/10.1093/ijpor/edq022

Harkness, J. A. (2003). Questionnaire translation. In J. A. Harkness, F. J. R. Van de Vijver, & P. Ph. Mohler (Eds.), *Cross-cultural survey methods* (pp. 35–36). Hoboken, NJ: Wiley-Interscience.

Harkness, J., Pennell, B.-E., & Schoua-Glusberg, A. (2004). Survey questionnaire translation and assessment. In S. Presser, J. Rothgeb, M. Couper, J. Lessler, E. Martin, J. Martin, & E. Singer (Eds.), *Methods for testing and evaluating survey questionnaires* (pp. 453–473). Hoboken, NJ: John Wiley & Sons.

Harkness, J. A., Villar, A., & Edwards, B. (2010). Translation, adaptation, and design. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, P. Ph. Mohler, B.-E. Pennell, & T. W. Smith (Eds.), *Survey methods in multinational, multiregional, and multicultural contexts* (pp. 115–140). Hoboken, NJ: John Wiley & Sons.

Holmes, J. (2008). *An introduction to sociolinguistic*s (3rd ed.). London, UK: Longman.

Hopkins, D., & King, G. (2010). Improving anchoring vignettes: Designing surveys to correct interpersonal incomparability. *Public Opinion Quarterly*, *74*, 201–222.

Hu, M., Lee, S., & Xu, H. (2018). Using anchoring vignettes to correct for differential response scale usage in 3MC surveys. In T. P. Johnson, B.-E. Pennell, I. A. L. Stoop, & B. Dorer (Eds.), *Advances in comparative survey methods: Multinational, multiregional, and multicultural contexts (3MC)*, (pp. 181–202). Hoboken, NJ: Wiley.

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. https://doi.org/10.1177/0022022189203004

Hunt, S. M., & Bhopal, R. (2004). Self report in clinical and epidemiological studies with non-English speakers: The challenge of language and culture. *Journal of Epidemiology and Community Health*, *58*(7), 618–622. https://doi.org/10.1136/jech.2003.010074

Institute of Medicine. (2003). *Unequal treatment: Confronting racial and ethnic disparities in health care*. Washington, DC: National Academies Press. https://doi.org/10.17226/12875

Javaras, K. N., & Ripley, B. D. (2007). An “unfolding” latent variable model for Likert attitude data. *Journal of the American Statistical Association*, *102*(478), 454–463. https://doi.org/10.1198/016214506000000960

Johnson, T. P. (1998). Approaches to equivalence in cross-cultural and cross-national survey research. In J. A. Harkness (Ed.), *Cross-cultural survey equivalence. ZUMA-Nachrichten Spezial 3* (pp. 1–40). Mannheim, Germany: ZUMA. Retrieved from https://www.ssoar.info/ssoar/handle/document/49730

Kandula, N. R., Lauderdale, D. S., & Baker, D. W. (2007). Differences in self-reported health among Asians, Latinos, and non-Hispanic whites: The role of language and nativity. *Annals of Epidemiology*, *17*(3), 191–198.

Keller, S. D., Ware, J. E., Gandek, B., Aaronson, N. K., Alonso, J., Apolone, G., … Wood-Dauphinee, S. (1998). Testing the equivalence of translations of widely used response choice labels: Results from the IQOLA Project. International Quality of Life Assessment. *Journal of Clinical Epidemiology*, *51*(11), 933–944.

Korey, J. L., & Lascher, E. L. (2006). Macropartisanship in California. *Public Opinion Quarterly*, *70*(1), 48–65. https://doi.org/10.1093/poq/nfj011

Lee, S., Nguyen, H. A., Jawad, M., & Kurata, J. (2008). Linguistic minorities in a health survey. *Public Opinion Quarterly*, *72*(3) 470–486. https://doi.org/10.1093/poq/nfn036

Marian, V., & Kaushanskaya, M. (2004). Self-construal and emotion in bicultural bilinguals. *Journal of Memory and Language*, *51*(2), 190–201. https://doi.org/10.1016/j.jml.2004.04.003

Marín, G., Gamba, R. J., & Marín, B. V. (1992). Extreme response style and acquiescence among Hispanics. *Journal of Cross-Cultural Psychology*, *23*(4), 498–509. https://doi.org/10.1177/0022022192234006

Marín, G., & Marín, B. V. (1991). *Research with Hispanic populations*. Thousand Oaks, CA: Sage Publications.

Mohler, P. P., Smith, T. W., & Harkness, J. A. (1998). Respondents' ratings of expressions from response scales: A two-country, two-language investigation on equivalence and translation. In J. A. Harkness (Ed.), *ZUMA-Nachrichten Spezial 3* (pp. 159–184). Mannheim, Germany: ZUMA.

Pennell, B.-E., Bowers, A., Carr, D., Chardoul, S., Cheung, G.-Q., Dinkelmann, K., … Torres, M. (2004). The development and implementation of the National Comorbidity Survey Replication, the National Survey of American Life, and the National Latino and Asian American Survey. *International Journal of Methods in Psychiatric Research*, *13*(4), 241–269.

Ross, M., Xun, W. Q. E., & Wilson, A. E. (2002). Language and the bicultural self. *Personality and Social Psychology Bulletin*, *28*(8), 1040–1050. https://doi.org/10.1177/01461672022811003

Ryan, C. (2013). *Language use in the United States: 2011*. Washington, DC. Retrieved from https://www2.census.gov/library/publications/2013/acs/acs-22/acs-22.pdf

Saris, W. E. (2003). Multitrait-multimethod studies. In J. A. Harkness, F. Van de Vijver, & P. Ph. Mohler (Eds.), *Cross-cultural survey methods* (pp. 207–233). Hoboken, NJ: Wiley-Interscience.

Scheuch, E. K. (1968). The cross-cultural use of sample surveys: Problems of comparability. *Historical social research/Historische sozialforschung*, *18*(2), 104–138.

Schwarz, N., Knauper, B., Hippler, H.-J., Noelle-Neumann, E., & Clark, L. (1991). Rating scales: Numeric values may change the meaning of scale labels. *Public Opinion Quarterly*, *55*(4), 570–582. https://doi.org/10.1086/269282

Shishido, K., Iwai, N., & Yasuda, T. (2009). Designing response categories of agreement scales for cross-national surveys in East Asia: The approach of the Japanese General Social Surveys. *International Journal of Japanese Sociology*, *18*(1), 97–111. https://doi.org/10.1111/j.1475-6781.2009.01111.x

Siegel, P., Martin, E. A., Bruno, R., Martin, E., & Siegel, P. (2001). Language use and linguistic isolation: Historical data and methodological issues. In *Statistical policy working paper 32: 2000 Seminar on integrating federal statistical information and processes* (Vol. 32, pp. 167–190). Washington, DC: Federal Committee on Statistical Methodology, Office of Management and Budget. Retrieved from https://www.census.gov/srd/papers/pdf/ssm2007-02.pdf

Smith, T. W. (2004). Developing and evaluating cross-national survey instruments. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, & E. Singer (Eds.), *Methods for testing and evaluating survey questionnaires* (pp. 431–452). Hoboken, NJ: John Wiley & Sons.

Smith, T. W. (2009). Editorial: Comparative survey research. *International Journal of Public Opinion Research*, *21*(3), 267–270. https://doi.org/10.1093/ijpor/edp038

Takeuchi, D. T., Gong, F., & Gee, G. (2012). The NLAAS Story: Some reflections, some insights. A commentary. *Asian American Journal of Psychology, 3*(2). https://doi.org/10.1037/a0029019

Toro-Morn, M. I. (2012). Familismo. In S. Loue & M. Sajatovic (Eds.), *Encyclopedia of immigrant health* (pp. 672–674). New York, NY: Springer Science + Business Media.

Tourangeau, R. (2004). Experimental design considerations for testing and evaluating questionnaires. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, & E. Singer (Eds.), *Methods for testing and evaluating survey questionnaires* (pp. 209–224). Hoboken, NJ: John Wiley & Sons. https://doi.org/10.1002/0471654728.ch11

Trafimow, D., Silverman, E. S., Fan, R. M.-T., & Fun Law, J. S. (1997). The effects of language and priming on the relative accessibility of the private self and the collective self. *Journal of Cross-Cultural Psychology*, *28*(1), 107–123. https://doi.org/10.1177/0022022197281007

Triandis, H. C., Davis, E. E., Vassiliou, V., & Nassiakou, M. (1965). *Some methodological problems concerning research on negotiations between monolinguals*. Urbana, IL: Department of Psychology, University of Illinois.

Van de Vijver, F. (2003). Bias and equivalence: Cross-cultural perspectives. In J. A. Harkness, F. Van de Vijver, & P. Ph. Mohler (Eds.), *Cross-cultural survey methods* (pp. 143–155). Hoboken, NJ: Wiley-Interscience.

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

Van Soest, A., Delaney, L., Harmon, C., Kapteyn, A., & Smith, J. P. (2011). Validating the use of anchoring vignettes for the correction of response scale differences in subjective questions. *Journal of the Royal Statistical Society. Series A, (Statistics in Society)*, *174*(3), 575–595. https://doi.org/10.1111/j.1467-985X.2011.00694.x

Viruell-Fuentes, E. A., Morenoff, J. D., Williams, D. R., & House, J. S. (2011). Language of interview, self-rated health, and the other Latino health puzzle. *American Journal of Public Health*, *101*(7), 1306–1313. https://doi.org/10.2105/AJPH.2009.175455

Weech-Maldonado, R., Elliott, M. N., Oluwole, A., Schiller, K. C., & Hays, R. D. (2008). Survey response style and differential use of CAHPS rating scales by Hispanics. *Medical Care*, *46*(9), 963–968. https://doi.org/10.1097/MLR.0b013e3181791924

Willis, G. B., Kudela, M. S., Levin, K., Norberg, A., Stark, D. S., Forsyth, B. H., … Hartman, A. M. (2010). Evaluation of a multistep survey translation process. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, P. Ph. Mohler, B.-E. Pennell, & T. W. Smith (Eds.), *Survey methods in multinational, multiregional, and multicultural contexts* (pp. 141–156). Hoboken, NJ: John Wiley & Sons.

Yan, T., & Hu, M. (2018). Examining translation and respondents' use of response scales in 3MC surveys. In T. P. Johnson, B.-E. Pennell, I. A. L. Stoop, & B. Dorer (Eds.), *Advances in comparative survey methods: Multinational, multiregional, and multicultural contexts (3MC)* (pp. 501–518). Hoboken, NJ: John Wiley & Sons.

Yu, S. M., Nyman, R. M., Kogan, M. D., Huang, Z. J., & Schwalberg, R. H. (2004). Parent's language of interview and access to care for children with special health care needs. *Ambulatory Pediatrics: The Official Journal of the Ambulatory Pediatric Association*, *4*(2), 181–187. https://doi.org/10.1367/A03-094R.1

Zea, M. C., Quezada, T., & Belgrave, F. (1993). Latino culture values: Their role in adjustment to disability. *Journal of Social Behavior and Personality*, *9*(5), 185–200.

Appendix 4-1. Question Names and Exact Wording

A. Excellent-to-Poor Scale

	English	Spanish
Scale	***Poor, Fair, Good, Excellent***	***Pobre, Regular, Bien, Excelente***
LP5a	How well do you speak Spanish?	¿Qué tan bien habla usted el español?
LP5b	How well do you read Spanish?	¿Qué tan bien lee usted el español?
LP5c	How well do you write in Spanish?	¿Qué tan bien escribe usted el español?
LP5d	How well do you speak English?	¿Qué tan bien habla usted el inglés?
LP5e	How well do you read English?	¿Qué tan bien lee usted el inglés?
LP5f	How well do you write in English?	¿Qué tan bien escribe usted el inglés?

B. Agreement Scale

	English	Spanish
Scale	***Strongly Agree, Somewhat Agree, Somewhat Disagree, Strongly Disagree***	***Mayormente de Acuerdo, Algo de Acuerdo, Algo en Desacuerdo, Mayormente en Desacuerdo***
FC Lead	Now I'd like to know how strongly you agree or disagree with the following statements about your family.	Ahora me gustaría saber qué tan de acuerdo o desacuerdo está con las siguientes descripciones sobre su familia.
FC1	Family members respect one another.	Los miembros de la familia se respetan unos a otros.
FC2	We share similar values and beliefs as a family.	Compartimos valores y creencias en común como familia.
FC3	Things work well for us as a family.	Las cosas resultan bien para nosotros como familia.
FC4	We really do trust and confide in each other.	Realmente compartimos y confiamos unos en otros.
FC5	Family members feel loyal to the family.	Sentimos mucha lealtad entre nosotros como familia.
FC6	We are proud of our family.	Estamos orgullosos de nuestra familia.
FC7	We can express our feelings with our family.	Podemos expresar nuestros sentimientos con nuestra familia.
FC8	Family members like to spend free time with each other.	A los miembros de la familia les gusta compartir el tiempo libre los unos con los otros.
FC9	Family members feel very close to each other.	Los miembros de la familia se sienten bien cercanos los unos de otros.
FC10	Family togetherness is very important.	La unión familiar es muy importante.

C. Frequency Scale

	English	**Spanish**
Scale	***Often, Sometimes, Rarely, Never***	***Muchas Veces, Alguna Veces, Casi Nunca, Nunca***
SN4	How often do your relatives or children make too many demands on you?	¿Con qué frecuencia exigen sus familiares demasiado de usted?
SN5	How often do your family or relatives argue with you?	¿Con qué frecuencia discuten o argumentan sus familiares con usted?
SN9	How often do your friends make too many demands on you?	¿Con qué frecuencia sus amigos(as) exigen demasiado de usted?
SN10	How often do your friends argue with you?	¿Con qué frecuencia discuten o argumentan sus amigos(as) con usted?

Scale	***Often, Sometimes, Rarely, Never***	***Muchas Veces, Alguna Veces, Pocas Veces, Nunca***
MS1_1	How often have you returned to [the country of origin of your parents/your country of origin]?	¿Con qué frecuencia ha regresado [the country of origin of your parents/your country of origin]?
DS4	How often do people dislike you because you are [ethnic/race group]?	¿Con qué frecuencia no le cae bien a la gente por ser de origen [ethnic/race group]?
DS5	How often do people treat you unfairly because you are [ethnic/race group]?	¿Con qué frecuencia le tratan injustamente por ser de origen [ethnic/race group]?
DS6	How often have you seen friends treated unfairly because they are [ethnic/race groups]?	¿Con qué frecuencia ha visto como tratan injustamente a sus amigos(as) por ser de origen [ethnic/race group]?

D. Quantity Scale

	English	**Spanish**
Scale	***A lot, Some, A little, Not at All***	***Mucho, Algo, Poco, Nada***
DA42 lead	As a result of the attacks, how much has your life been affected in the following areas –?	Debido a los ataques de terrorismo, ¿cuánto se ha visto afectada su vida en las siguientes áreas?
DA42b	Losing my job.	Perder mi trabajo.
DA42d	Reduction in my family income.	Tener una reducción en el ingreso familiar.
DA42e	Feeling more patriotic.	Sentirme más patriótico(a).
DA42f	Feeling less safe and secure.	Sentirme menos a salvo e inseguro(a).
DA42h	Been treated unfairly because of my race, ethnicity, or physical appearance.	Tener un trato injusto por mi raza, origen étnico, o apariencia física.
DA42i	Feeling less optimistic about the future.	Sentirme menos optimista acerca del futuro.
DA42l	Feeling that I no longer can cope with things.	Sentirme que no puedo hacerle frente a las cosas.

Scale	***A Lot, Some, A Little, Not at All***	***Mucho, Regular, Un Poco, Nada***
SN2	[Not including your husband/wife/partner] how much can you rely on relatives who do not live with you for help if you have a serious problem?	[Sin incluir a su esposo/esposa/pareja] ¿cuánto puede contar con que los familiares que no viven con usted lo (la) ayuden si tiene un problema serio?
SN3	[Not including your husband/wife/partner] how much can you open up to relatives who do not live with you if you need to talk about your worries?	[Sin incluir a su esposo/esposa/pareja] ¿cuánta confianza puede tener con los familiares que no viven con usted si necesita hablar de sus preocupaciones?
SN7	How much can you rely on your friends for help if you have a serious problem?	¿Cuánto puede contar con que sus amigos(as) lo (la) ayuden si tiene un problema serio?
SN8	How much can you open up to your friends if you need to talk about your worries?	¿Cuánta confianza tiene usted con sus amigos(as) si necesita hablar de sus preocupaciones?

Appendix 4-2. Coefficient Estimates of ANOVA for All Measures

(Bold indicates significant at *p* < .1)

A. Excellent-to-Poor Scale

	LP5a	LP5b	LP5c	LP5d	LP5e	LP5f
Intercept	**1.872**	**1.097**	**0.725**	**1.778**	**1.788**	**1.535**
Language: English vs. Spanish	0.078	–0.019	–0.166	0.119	**0.146**	0.072
Score summary: Total score quartiles	**0.504**	**0.720**	**0.790**	**0.556**	**0.576**	**0.633**
Language × score summary	–0.006	0.021	0.043	–0.026	**–0.051**	–0.033

B. Agreement Scale

	FC1	FC2	FC3	FC4	FC5	FC6	FC7	FC8	FC9	FC10
Intercept	0.180	–0.082	0.054	**–0.260**	–0.077	0.169	**–0.292**	**–0.405**	–0.162	0.224
Language: English vs. Spanish	–0.214	0.066	–0.242	–0.108	–0.217	–0.079	0.126	0.181	–0.242	–0.140
Score summary: Total score quartiles	**0.435**	**0.544**	**0.474**	**0.602**	**0.527**	**0.396**	**0.631**	**0.700**	**0.581**	**0.366**
Language × score summary	0.084	–0.018	**0.127**	0.056	0.080	0.027	–0.053	–0.049	0.078	0.071

C. Frequency Scale

	SN4	SN5	SN9	SN10
Intercept	**0.995**	**1.940**	**1.797**	**2.242**
Language: English vs. Spanish	–0.074	**–0.407**	**0.562**	**0.501**
Score summary: Total score quartiles	**0.673**	**0.417**	**0.529**	**0.425**
Language × score summary	0.006	**0.118**	**–0.160**	**–0.150**
	MS1_1	**DS4**	**DS5**	**DS6**
Intercept	**1.390**	**1.994**	**1.888**	**1.468**
Language: English vs. Spanish	–0.055	0.106	0.204	–0.089
Score summary: Total score quartiles	**0.469**	**0.469**	**0.540**	**0.590**
Language × score summary	0.039	–0.007	–0.051	0.023

D. Quantity Scale

	SN2	SN3	SN7	SN8
Intercept	**0.531**	**0.521**	**0.509**	**0.284**
Language: English vs. Spanish	–0.182	–0.159	0.086	0.142
Score summary: Total score quartiles	**0.464**	**0.473**	**0.667**	**0.712**
Language × score summary	0.091	**0.158**	–0.081	–0.070

	DA42b	DA42d	DA42e	DA42f	DA42h	DA42i	DA42l
Intercept	**2.568**	**2.214**	**0.722**	**0.837**	**3.223**	**1.883**	**3.422**
Language: English vs. Spanish	**–0.372**	–0.237	0.137	0.177	0.046	–0.133	–0.117
Score summary: Total score quartiles	**0.363**	**0.468**	**0.497**	**0.692**	**0.188**	**0.516**	**0.159**
Language × score summary	0.114	0.071	–0.008	–0.088	0.001	0.034	0.028