Language Barriers: Causal Evidence of Linguistic Item Bias in Multilingual Surveys

Yamil Ricardo Velez* Angel Saavedra† Jose Gomez-Espinal‡

December 5, 2021

Abstract

Accurate estimation of public opinion in linguistically diverse societies depends on survey questions that operate similarly across groups. Though scholars have documented the presence of item bias, also known as differential item functioning (DIF), in multilingual surveys, extant studies have used observational designs, complicating whether DIF can be attributed to survey language or unobserved differences between linguistic groups. Moreover, it is unclear whether linguistic DIF is due to translation errors or the psychological effects of language. We leverage 1,953 English-Spanish bilinguals across three studies to assess how survey language affects item properties. We find evidence of DIF across a variety of items measuring identity, attitudes, and political knowledge. Perceived translation quality does not predict DIF and items with minimal text still exhibit different item properties, suggesting that factors other than translation quality can contribute to the problem. These findings raise important questions about the design of multilingual surveys.

*Yamil Ricardo Velez is an assistant professor in the Department of Political Science at Columbia University, 420 W. 118th Street, New York, NY, 10027 (yrv2004@columbia.edu).

†Angel Saavedra is an assistant professor in the Department of Political Science at St. Norbert College, 100 Grant Street, De Pere, WI, 54115 (angel.saavedracisneros@snc.edu).

‡Jose Gomez-Espinal is a Ph.D. student in the Department of Political Science at Columbia University, 420 W. 118th Street, New York, NY, 10027 (jg3341@columbia.edu)

Introduction

Scholars have drawn attention to the economic and political integration of minorities across the globe; groups that also include linguistic minorities (Ramakrishnan and Ahmad 2014). Multilingual surveys are necessary to describe their preferences and behavior, but designing them can be difficult due to differences in meaning across languages (Iyengar 1993). Survey researchers often hire expert translators or assemble diverse teams of raters to evaluate the consistency of words across dialects, using the process of back-translation to guarantee that translated items possess similar meanings when translated back to their original language (Brislin 1970). However, even the highest-quality translation cannot ensure that items will be interpreted similarly by respondents (Sireci 1997, 12).

Scholars have accumulated evidence demonstrating gaps in attitudes between English and Spanish-speaking Latino interviewees (Welch, Comer, and Steinman 1973; Lee and Pérez 2014; Hill and Moreno 2001). Researchers have also documented similar differences among Asian Americans (Lien, Conway, and Wong 2003). Though these gaps could reflect "true" differences between linguistic groups, they could also be due to measurement artifacts. Using measurement models, scholars have found that survey items are not functionally equivalent across language forms (Pérez 2009, 2011).

Though prior studies suggest that some measures are not functionally equivalent (Pérez 2009), it is possible that unobserved differences between linguistic groups are responsible. Existing studies adjust for variables that covary with survey language selection. However, unobserved variables could still affect tests of item bias (also known as differential item functioning). As a result, translated items could possess lackluster measurement properties because of pre-existing differences between English and non-English speakers (Sireci 1997). For instance, an item measuring political knowledge might operate differently in English and Spanish not because translations are poor, but because English and Spanish-dominant respondents are embedded in different social networks, and thus, possess different understandings of politics. A downstream consequence of this is that scores on attitudinal or knowledge scales will differ across language forms due to respondent characteristics, rather than aspects of the survey itself.

We make two contributions in this paper. First, we conduct an analysis of differential item functioning using bilingual respondents who are randomly assigned to language forms. This design ensures that DIF can be attributed to survey language. Using this design, we find consistent evidence of DIF across scales measuring identity, attitudes, and policy preferences. Second, we document the possibility of DIF due to the psychological effects of language, rather than mere translation errors. First, we use expert ratings of translation quality to assess if there is a relationship between quality and DIF. We fail to find evidence of any such relationship. We also minimize the impact of translation by using a novel visual item format and still detect DIF. This suggests that evidence of linguistic DIF need not signify poor translations; they could also be the result of the differential accessibility and salience of concepts across languages. Our findings highlight the need to pre-test multilingual items, given that high-quality translations are unlikely to address this type of DIF. They also speak to general challenges in measuring opinion for linguistically diverse groups.

The Challenges of Survey Translation

Multilingual surveys are informative when translated terms communicate similar meanings. Methods such as back-translation aim to achieve this goal by translating a survey from a source to a target language, translating the survey back to the source language, and minimizing discrepancies between the original and back-translated questionnaire (Brislin 1970). However, even when survey forms are back translated and questions are deemed to be literally equivalent, words can still possess different valences (Ervin and Bower 1952) and meanings (Pérez 2009) that escape the notice of translators and survey researchers.

Following research in psychometrics, political scientists have devoted attention to the general problem of DIF. Differences between racial and ethnic groups in political knowledge have been shown to be a function of "perceptual biases" that alter how voters of color interpret political knowledge questions (Abrajano 2015). Common scales used in the American National Election Study such as egalitarianism and support for limited government suffer from DIF, such that racial and ethnic minorities scoring at similar levels as whites on a variety of scales respond differently to certain items (Pietryka and Macintosh 2017). Language-of-interview differences have been detected among Latinos, even after adjusting for demographic covariates (Lee and Pérez 2014). Linguistic DIF has also been observed using other methods such as structural equation modeling and multiple indicators multiple causes (MIMIC) models (Pérez 2009, 2011).

Extant studies have been carried out in observational settings where unobserved confounding may threaten inferences. For example, linguistic DIF might be due to unmeasured factors such as geography or social networks. Latinos embedded in politically active social networks might be more likely to score higher on scales measuring identity strength or political knowledge, and this might covary with language usage. Since the direction and magnitude of bias due to unobserved confounding is difficult to ascertain, it remains unclear whether different survey forms or unobserved factors influencing survey language selection are responsible for DIF. Observing evidence of linguistic DIF might imply a focus on improving survey translations, whereas DIF due to a factor such as geography, for example, might entail developing questions that are relevant across regions. Therefore, distinguishing between these alternative explanations is important for improving multilingual surveys.

Minimizing unobserved confounding in the case of DIF detection requires a random assignment mechanism and a population that is capable of completing surveys in multiple languages. In recent years, studies have leveraged bilinguals in experimental settings to isolate the effects of language (or linguistic features) on political attitudes (Pérez and Tavits 2016, 2017). By virtue of random assignment, this design ensures the bilingual respondents are balanced on pre-treatment covariates (e.g., age, education, citizenship), and allows us to recover the effect of assignment to a language form on subsequent attitudes. Though these recent studies have illustrated the advantages of using bilinguals to assess the effects of language on *attitudinal scales*, this design has yet to be applied to the question of DIF in political science. Indeed, Pérez (2011, 448) recommends experimental designs using bilingual populations as a way of circumventing the issue of unobserved confounding when evaluating the presence of linguistic DIF; a recommendation that is echoed by psychometrics research on test translation (Sireci 1997; Sireci and Berberoglu 2000). Sireci and Berberoglu (2000) randomly assign Turkish-English bilingual students to one of two course evaluation forms that present Turkish and English items in alternating order, and find significant evidence of DIF in items measuring ambiguous concepts. Our study follows these recommendations to evaluate if DIF can be detected in settings where unobserved confounding can be addressed via randomization.

Data and Methods

We investigate linguistic DIF using three diverse samples of bilingual Latinos. The first sample is comprised of students at a Hispanic Serving Institution (HSI) in the Southwestern United States. The survey was in the field from August 9, 2017 to December 13, 2017.[1] In total, 194 students participated in the study. The second sample is derived from a sample of Latinos collected on the Lucid platform from November 18, 2019 to November 26, 2019. Our survey was only accessible to respondents who self-identified as Latino. 1,520 Latinos completed the survey, 1,261 of whom were bilingual. Finally, the third sample is based on a Latino sample recruited by CloudResearch. Much like Lucid, CloudResearch relies on a marketplace model encompassing many sample providers. The survey was in the field from October 30, 2020 to November 15, 2020 and yielded 1,918 respondents, 1,552 of whom were bilingual. Though these are convenience samples, Appendix B shows comparability to national samples of Latinos on dimensions such as age, ideology, and national origin. Moreover, given the growing

[1] Students were recruited to participate in a study on Latino/a politics via instructor e-mails. A $50 Amazon.com gift card was used to incentivize participation.

use of these online panels, identifying DIF in these settings is important.

Study 1 featured scales that have been included in previous work (i.e., Latino attachment, beliefs about Americanism, immigration opinion) (Pérez 2009, 2011). We also included a measure of partisanship as a social identity to assess if political identities were more or less susceptible to DIF (Huddy, Mason, and Aarøe 2015). These English and Spanish scales, with the exception of the partisan identity scale, were all taken from the 2006 Latino National Survey (LNS). The scales for the LNS were professionally translated. The partisan identity scale was translated by the authors. Study 2 included measures of Latino identity and immigration policy attitudes.[2] Translated items were drawn from a mixture of author-translated items, the 2006 LNS and the Hispanic oversample of the 2016 ANES Time Series study. These items were slightly adapted to include new identity terms such as "Latinx." Study 3 measured policy preferences, identity, political knowledge, preferences for minority representation, and political efficacy. These items were drawn from a mixture of professional and author translations. Question wording for these items is reproduced in Appendix A.

In all of the studies, bilingual respondents were randomly assigned to one of three forms: (1) a Spanish form, (2) an English form, or (3) a hybrid form that exposed participants to random English and Spanish-language subsets of each scale. Those assigned to the Spanish (English) form responded to all items in Spanish (English), whereas those assigned to the hybrid form were randomly assigned to English and Spanish items for each scale. For instance, for a four-item scale, respondents were first assigned to two items in English and then two unseen items in Spanish.[3] This was done to eliminate consistency effects associated with seeing the same item twice. One advantage of the hybrid form is that items can be placed on a common metric. Other studies assume the presence of anchor items, or items that operate identically across

[2]Upon being assigned to a language form, respondents were also randomly assigned to different prompts asking them to write about the importance of their ethnic identity. These treatments had no discernible effects on attitudes. We use the complete sample to preserve statistical power.

[3]The first study randomized the order of languages. However, to simplify survey programming, the subsequent studies first assign English items before Spanish items. We assume that there are no order effects.

languages, but this assumption is never verified (Pérez 2009).

Our sample for the HSI study indicated a high degree of bilingualism, with over 90% of students reporting speaking English *and* Spanish "well" or "very well." All subjects in this study were randomly assigned to one of the three forms. In Studies 2 and 3, we included a two-question pre-treatment language quiz (Flores and Coppock 2018), and randomly assigned self-reported bilinguals to one of the three forms. For our analyses, we subset on respondents who passed the two-question language quiz to ensure that respondents who exhibit at least a basic proficiency are compared. This also indirectly serves as an attention check, given that participants had to read two brief vignettes in order to respond correctly. This yields final sample sizes of 840 for the Lucid study and 919 for the CloudResearch study. We replicate the key DIF findings using the full sample of bilinguals and present them in Appendix C.

Though there are various ways to assess DIF such as the Mantel-Haenszel method, MIMIC models, and logistic regression, we rely on item response theory (IRT) models, given our focus on understanding how item properties relate to survey language.[4] We estimate outcome-appropriate two-parameter IRT models, represented by the general equation below:[5]

$$Y_{ij} = f(\theta_i, \alpha_{jl}, \beta_{jl}) \tag{1}$$

where i represents individuals, j represents items, θ represents the latent trait or attitude, l denotes the item language, α represents the item difficulty parameter, and β represents the item discrimination parameter.[6] Item discrimination captures the de-

[4]Studies comparing the DIF method presented here to other approaches such as MIMIC have found that MIMIC models tend to have better control over Type I error rates for longer batteries (50 items or more) that are more common in educational testing contexts (Finch 2005). The goal of this paper is not to adjudicate between these different methods, but to show that linguistic DIF can be detected in an experimental setting using a popular DIF detection method.

[5]In Appendix E, we estimate one-parameter versions of each IRT model presented in the paper, and still detect evidence of DIF for most scales.

[6]For the HSI study, we estimate simpler one-parameter models to avoid over-fitting, given the smaller sample size (N = 194). These models estimate a difficulty parameter for each binary item and item step difficulties for each polytomous item. Item step difficulties capture the point on the latent scale where probability curves for two adjacent response categories intersect (e.g., the point on the latent scale at which the probability of "somewhat agree" is equal to "strongly

gree to which an item distinguishes between those who score low and high on the latent scale, whereas difficulty corresponds to where an item falls on the latent scale. Likelihood-ratio (LR) tests provide a method for detecting DIF. First, two sets of models are estimated: a constrained model that fixes item parameters to be equal across languages (e.g., $\alpha_{j0} = \alpha_{j1}$) and an unconstrained model that allows item parameters to be freely estimated. These models are then compared using fit statistics:

$$G^2_{df} = [-2log\ likelihood_{constrained}] - [-2log\ likelihood_{unconstrained}] \tag{2}$$

with degrees of freedom (df) equal to the difference in parameters between the constrained and unconstrained model. The null hypothesis is that a constrained model assuming no DIF explains the data as well as a model that allows item parameters to vary by language. Thus, rejecting the null hypothesis can be considered evidence of DIF. The LR test has been shown to control Type I error rates (Ankenmann, Witt, and Dunbar 1999; Atar and Kamata 2011). We estimate multiple-group IRT models using the *TAM* package in *R* and fix the means of θ across language forms to zero.[7] Estimation is carried out using marginal maximum likelihood.

Results

We first describe the results of our omnibus tests of DIF before characterizing the prevalence and direction of DIF. As shown in Table 1, we reject the null of measurement equivalence in 9 out of 12 cases. In the HSI sample, we reject the null hypothesis of no DIF for the Latino attachment ($G^2_{(8)}$ = 25, p<.001) and immigration opinion ($G^2_{(12)}$ = 268, p<.001) scales. For the Americanism beliefs and partisan identity scales, we fail to reject the null hypothesis. Therefore, the hypothesis that the English and Spanish items are statistically equivalent is rejected in the case of Latino attachment and immigration opinion, but not Americanism beliefs and partisan identity. In the Lucid sample, we reject the null hypothesis of no DIF for all three scales. The DIF model yields

agree") (Nering and Ostini 2011).

[7] Given the randomization of participants to language forms, we view this as a plausible assumption.

statistically significant improvement in fit to the data relative to the no DIF model in the case of panethnic identity ($G^2_{(10)}$ = 24, p<.001), anti-trans attitudes ($G^2_{(42)}$ = 78, p<.001), and immigration opinion ($G^2_{(18)}$ = 52, p<.001). Finally, in the CloudResearch sample, we also reject the null hypothesis of no DIF for all of the available scales but one (minority representation). The DIF model provides an improvement in fit relative to the no DIF model in the case of Latino attachment[8] ($G^2_{(17)}$ = 92, p<.001), ideology ($G^2_{(13)}$ = 27, p=.01), political efficacy ($G^2_{(24)}$ = 413, p<.001), and political knowledge ($G^2_{(14)}$ = 54, p<.001). In sum, the DIF-afflicted scales span constructs such as identity, policy preferences, ideology, political knowledge, and general political beliefs.

Table 1: Likelihood-ratio tests of DIF

Sample	Translation Source	Outcome	Model	G^2	df	p
HSI	2006 LNS	Latino Attachment	PCM	25.25	8	0.00
HSI	2006 LNS	Americanism Beliefs	PCM	10.91	8	0.21
HSI	Authors	Partisan Identity	PCM	15.48	10	0.12
HSI	2006 LNS	Immigration Opinion	NRM	268.41	12	0.00
Lucid	2016 ANES/Authors	Panethnic Identity	GPCM	24.64	10	0.01
Lucid	Authors	Trans Attitudes	GPCM	78.27	42	0.00
Lucid	2016 ANES/Authors	Immigration Opinion	GPCM	52.35	18	0.00
CR	2006 LNS/Authors	Latino Attachment	GPCM	91.79	17	0.00[9]
CR	2010/2016 Latino CES	Ideology	2PL	26.52	13	0.01
CR	Authors	Political Efficacy	GPCM	413.18	24	0.00
CR	Authors	Minority Representation	GPCM	20.89	15	0.14
CR	2010 Latino CES/Authors	Political Knowledge	2PL	54.15	14	0.00

Though the results above suggest the presence of DIF, the direction of DIF is unclear. That is, we do not know if items are consistently easier in English or Spanish. Knowing this is important as it could illuminate where changes need to be made to surveys so that they can be made more comparable. It could also speak to possible mechanisms. For instance, if items measuring attitudes about American politics are consistently easier in English whereas those capturing attitudes about Latin America are consistently easier in Spanish, this might suggest that linguistic DIF operates via differences in accessibility of concepts across languages. It is also important to note that the LR-test method assesses if DIF is present, but does not provide any indication of its prevalence within a scale. If a single item evinces DIF, one can remove this item

[8]This scale was identical to the HSI version, except for the inclusion of visual items.

and carry out estimation with the DIF-free scale, a process referred to as "test purification" (Hidalgo-Montesinos and Gómez-Benito 2003). However, if many items display evidence of DIF, a general lack of comparability might call the use of the entire scale into question.

We move beyond the scale-level analysis presented above, and assess differences in item difficulties within the DIF-afflicted scales.[10] For each item, we compute the difference between the English and Spanish item difficulties, along with the standard error of the difference, and code whether there is significant evidence of a difference between English and Spanish versions of the same item ($p<.05$; two-tailed). Given that models vary in the number of estimated parameters, we graphically depict these differences using a two-dimensional discrete heat-map. The horizontal axis represents an response category step (i.e., movement from a response option to the adjacent response option), whereas the vertical axis represents an item. For binary items, there is only one cell because respondents can only move from 0 to 1, whereas for ordinal items, the number of cells within each row represent the number of response category steps. Items that are easier in English are indicated in light gray, those that are easier in Spanish are indicated in dark gray, and those for which we possess insufficient evidence of DIF are indicated using a pale gray. Examining the entire set of scales, we see that DIF is very prevalent. The number of items afflicted by DIF ranges from one (immigration opinion in the HSI sample) to six (anti-trans attitudes in the Lucid sample).

To what extent does DIF operate in a single direction? Beyond showing that DIF is prevalent, Figure 1 reveals a considerable amount of heterogeneity within scales with respect to linguistic differences in item difficulties. There are 9 instances of DIF that render an item easier in Spanish across all item categories, 15 instances of DIF that render an item easier in English across all item categories, and 8 instances of DIF for which a subset of response categories is easier in Spanish and another subset is easier

[10] We present item discrimination differences in Appendix F. We focus on item difficulties because they are easier to interpret and more commonly used in substantive applications of IRT.

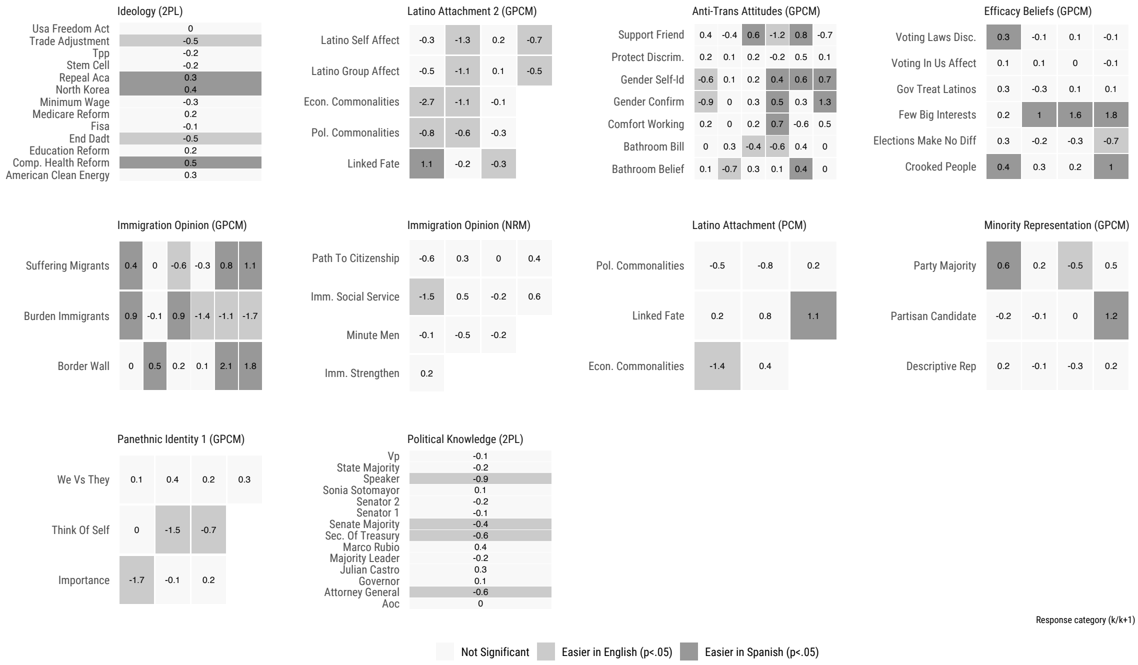

Figure 1: Differences in item difficulties across languages. For each scale, items are presented on the vertical axis and response category steps are on the horizontal axis. Estimated differences are reported inside each cell. Significant differences at the .05 level (two-tailed) suggestive of DIF that favors English (Spanish) items are denoted in light (dark) gray. Two-parameter logistic models (2PL), generalized partial credit models (GPCM), and nominal response models (NRM) are estimated for scales comprised of binary, ordinal, and categorical outcomes, respectively. Models are estimated in R using marginal maximum likelihood (TAM package).

in English.[11] In sum, the direction of DIF is difficult to predict, and this can even vary within items.

Mechanisms

Our bilingual design allows us to identify linguistic differences that arise due to the surveys respondents take, rather than the characteristics of survey takers. Through randomization, mean differences between participants assigned to language forms will be zero in expectation. This means that if a survey is poorly mistranslated, our experimental approach will likely detect it because items in the scale will possess different item properties. Poor translations do not call the causal status of our design into question; rather, translation quality can be considered a mechanism that explains the

[11]These "mixed" cases can be observed in the panethnic identity, anti-trans attitudes, and immigration opinion scales, where the selection of lower response options is easier in English, but the selection of higher response options is easier in Spanish, and vice versa.

mean difference in item properties across language forms.

A simple assessment of whether high-quality translations such as those present in the 2006 LNS and 2010/2016 Latino CES are less susceptible to DIF than those translated by the authors suggest that translation quality does not explain the pattern of results. Of the eight scales that included items featured in nationally representative samples of Latinos, 88% suffered from DIF. Of the four scales exclusively comprised of items featured in national samples, 75% suffered from DIF. Assuming that nationally representative surveys of Latinos employing professional translators possess "high translation quality," this pattern of results suggests that quality alone is unlikely to explain the DIF that is observed across scales and samples.

To take a more systematic look at the relationship between translation quality and DIF estimates, we recruited two expert university translators who were blind to the purpose of the study and not given any information about the translation source. Translators were presented with survey questions side-by-side in English and Spanish and asked to rate whether the translation was accurate (i.e., whether words were comparable across languages), grammatically correct, and not awkward or disfluent. 5-point Likert scales ranging from "strongly disagree" to "strongly agree" were used to measure these three components. A text box allowed the translators to report errors with imperfect survey questions. Translators took just over two hours on average to complete the translation assessment ($\bar{x}$ = 2h 21m). These three ratings were then aggregated to create an overall score of translation quality. This scale displayed moderate levels of reliability ($\alpha = .53$). These mean ratings are summarized for each scale in Table 2.

Though 91% of the items were rated above the midpoint by both translators, there is a weak correlation between the ratings across translators (r = .10). Overall, Translator 2 was a "harsher" rater than Translator 1. Compared to Translator 1, whose average rating was 4.5 for the entire set of scales, Translator 2's average rating was 3.97. While scales featured in national surveys of Latinos such as "Latino Attachment" and "Immigration Opinion" were rated on the lower end of translation quality by both translators, there were some discrepancies in ratings. For instance, the translation of the Kalla and Broockman (2020) scale of anti-trans sentiment was rated as being of high quality by Translator 1 ($\bar{x} = 5$), but was perceived to be of moderate quality by Translator 2 ($\bar{x} = 3.83$). Measures of panethnic and partisan identity were rated highly by Translator 1 ($\bar{x_1} = 5$; $\bar{x_2} = 5$), but were of mediocre quality according to Translator 2 ($\bar{x_1} = 3$; $\bar{x_2} = 3.89$). Finally, the ideology scale comprised of roll-call items that was featured on the Latino CES was rated highly by Translator 1 ($\bar{x} = 4.67$), but received a substantially lower score from Translator 2 ($\bar{x} = 3.33$). Given the high level of disagreement across translators, we evaluate if quality ratings are systematically related to linguistic DIF by including each translators' ratings separately in a single model predicting DIF estimates.[12]

Table 2: Expert ratings of translation quality for each scale

Scale	Translator 1	Translator 2
Immigration Opinion[a] (HSI)	2.92	3.67
Latino Attachment[a] (HSI/CR)	3.33	3.78
Americanism Beliefs[a] (HSI)	4.25	4.67
Immigration Opinion[b] (Lucid)	4.44	4.33
Ideology[a] (CR)	4.67	3.33
Efficacy Beliefs (CR)[b]	4.89	4.28
Political Knowledge[c] (CR)	4.90	4.60
Panethnic Identity[c] (Lucid)	5.00	3.00
Anti-Trans Attitude Scale[b] (Lucid)	5.00	3.83
Partisan Identity[b] (HSI)	5.00	3.89
Political Representation[b] (CR)	5.00	4.33

Source: (a) denotes a translated scale featured in a national sample of Latinos; (b) denotes an author-translated scale; and (c) represents a scale involving a mixture of author-translated and professionally translated scales. See Table 1 for full list.

Table 3 estimates the relationship between the two separate measures of translation quality and the mean absolute DIF for every survey question.[13] We compute the mean absolute DIF by taking the absolute value of the DIF estimates for each step (i.e., movement from one response category to a higher response category), summing them, and dividing them by the number of steps for a given survey item.[14] Translation qual-

[12]We also observe that ratings do not predict DIF estimates if the two ratings are averaged.
[13]Similar findings hold even if DIF is allowed to take on positive and negative values.
[14]For binary items, the number of steps is one (i.e., the probability of moving from 0 to 1),

Table 3: Linear model regressing mean absolute DIF estimates for each survey question on translation ratings

	Dependent variable:
	Mean Absolute DIF
Intercept	0.585* (0.299)
Translation Quality (Translator 1)	−0.014 (0.054)
Translation Quality (Translator 2)	−0.023 (0.049)
Observations	60
R^2	0.005
Residual Std. Error	0.289 (df = 57)
Note:	*p<0.1; **p<0.05; ***p<0.01

ity does not appear to predict DIF using the two expert ratings. The slope coefficients for the quality ratings are small, not statistically significant, and the R^2 for a model regressing DIF estimates on the two estimates of translation quality is .005. Thus, the evidence presented here suggests that translation quality does not seem to account for the wide variation in DIF estimates across survey items, indicating that factors other than translation may be responsible for DIF.

Though faulty translations are a natural culprit for DIF, recent research has indicated that linguistic differences in surveys can emerge because different languages facilitate the retrieval of distinct concepts from memory (Pérez and Tavits 2019). Put another way, the presence of linguistic DIF need not imply that there are errors in the survey translation process; it can also reflect a cognitive process of differential accessibility, whereby distinct political considerations are accessible depending on the survey language. As an example, Pérez and Tavits (2017) find that bilingual Russian-Estonians assigned to respond to surveys in Estonian (a future-less tongue) are less likely to support policies that mitigate climate change than those assigned to take the survey in Russian (a futured tongue). These general findings are consistent with a stream of research considering the impact of language on cognition, finding that lan-

whereas for ordinal items, the number of steps is equal to K-1, where K represents the number of response categories. For example, four step difficulties are estimated when one is modeling a five-point scale. These step difficulties capture the probability of selecting the second response category versus the first, the third versus the second, and so on.

guage can influence perceptions of time (Boroditsky 2001) and structure interpretations of events (Slobin 1996). Among bilinguals, languages can prime distinct value orientations and activate more intense emotional states when there is a match between the language spoken during the moment of encoding and retrieval (Marian and Kaushanskaya 2004).

We use data from our third study to investigate this alternative mechanism. Specifically, we evaluate if linguistic DIF can still be detected when the impact of translation is minimized by using items and response options with visual guides. Translation-based DIF can be attributed to either question wording and/or incomparable response options. Though DIF due to question wording is well-known, DIF might also operate through incomparable response options (King and Wand 2007). Holding question content fixed, it could be the case that response options possess different valences. For example, words like "very" and "muy" might be situated on different points of the latent space with respect to enthusiasm. If this is the case, altering the survey question posed to respondents is unlikely to resolve the DIF.

We design two sets of items. First, we construct political knowledge items with images of politicians and minimal text (see Figure 2). As Prior (2014) notes, this format can facilitate the retrieval of political concepts from memory among people who have a more visual cognitive style. However, the use of visual items has an additional strength, which is that it allows simpler survey questions to be constructed, thus minimizing the importance of translation. Second, we incorporate items with visual response options into two scales (panethnic identity and efficacy) and assess if we still observe linguistic DIF for these items (see Figure 3). These items also employ simple language (e.g., "how does being Latino make you feel?"). For both sets of items, our expectation is that if DIF is detected, translation error is a less probable explanation, given the limited presence of language in the question text and/or rating scales.

Figure 4 displays the difference in item difficulties between English and Spanish items measuring political knowledge. Negative scores indicate items that are easier in English. Positive scores indicate items that are easier in Spanish. Evidence of DIF is

Figure 2: Example of Political Knowledge Item with Visual Cue.

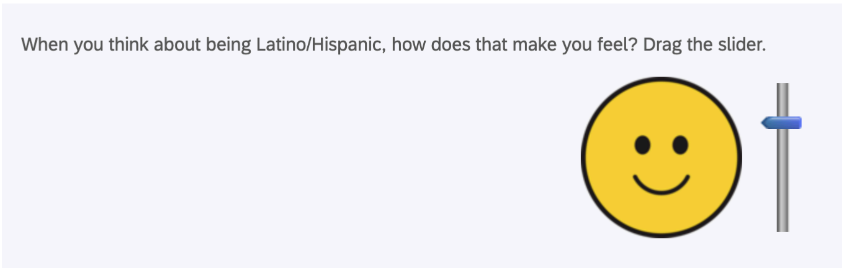

Figure 3: Example of Latino Attachment Item with Visual Response Options.

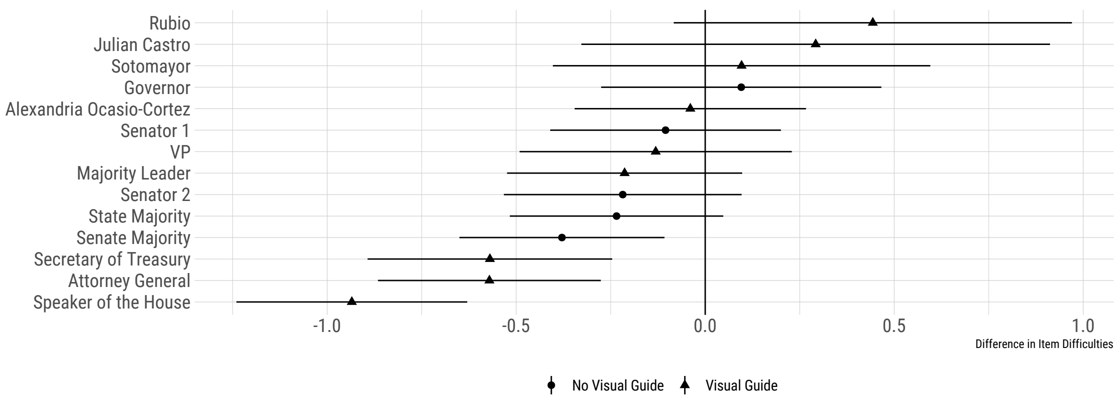

Figure 4: Differences in item difficulties and corresponding SEs.

detected in four out of 14 items, 3 of which involve pictorial items and possess lower item difficulties in English ($p < .05$; two-tailed). These differences in item parameters

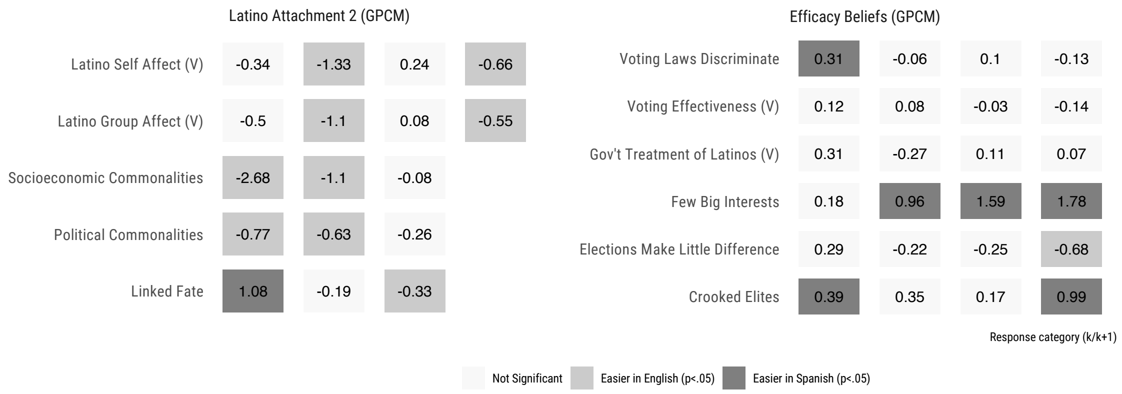

Figure 5: Differences in item difficulties across languages for the panethnic identity and efficacy beliefs scales. For each scale, items are presented on the vertical axis and response categories are on the horizontal axis. Estimated differences are reported inside each cell. Significant differences at the .05 level (two-tailed) suggestive of DIF that favors English (Spanish) items are denoted in light (dark) gray.

are sizable, ranging from .38 to .94 units on the latent logistic scale. In practice, this means that otherwise similar people may be classified as having different levels of latent political knowledge by virtue of the survey form they complete. Even when text is simplified and images are used, we detect evidence of DIF, which suggests that translation errors are unlikely to be responsible.

We now turn to our analysis of items with non-verbal response options. Recall, these items replace traditional Likert-style response options (e.g., “strongly agree”) with visual response options (i.e., faces). Panel 5 displays item parameter differences for the panethnic identity and efficacy beliefs scale; visual items are marked with a (V) next to the item name. As shown in the figure, visual items significantly evince DIF for the panethnic identity scale, but not for the efficacy beliefs scale. The two visual items ask respondents to indicate their level of positive (or negative) affect regarding their own Latino/Hispanic identity, as well as how they feel about other group members. In both cases, the item is easier in English than in Spanish; a pattern observed in other scale items (e.g., socioeconomic commonalities, political commonalities, linked fate). In contrast, there is no statistically discernible evidence of DIF for the non-verbal items in the efficacy beliefs scale, but evidence of DIF as it relates to verbal items measuring perceptions of elites and voting laws. Taken together, we find that DIF can be detected even when minimal text and visual response options are used, which suggests that a mechanism other than translation error is responsible.[15] This analysis, of course, does not rule out translation error as a reason for observing linguistic DIF. Instead, the purpose of these tests is to highlight an additional mechanism – differential accessibility – that might account for instances where translation errors are unlikely to be responsible.

Concluding remarks

Translation is becoming an increasingly important aspect of survey research in ethnically diverse societies. A distinct challenge facing multilingual surveys is that translations can render survey questions incomparable. Though previous studies have documented the possibility of linguistic DIF in multilingual surveys, extant work has been unable to parse out the unique effects of survey language. In this paper, we leveraged bilinguals to estimate the causal effect of language and uncovered evidence that even among similar survey respondents who are otherwise assigned to different survey forms, linguistic DIF was detected. These patterns were prevalent across scales, samples, and time. We also provided evidence that DIF could be detected even after the impact of translation was minimized. This suggests that not all DIF can be attributed to translation errors, and implies that even high-quality translations may be susceptible to DIF.

Though these analyses confirm that linguistic DIF can be recovered even in experimental settings, there are caveats worth mentioning. First, bilinguals might not be representative of their monolingual counterparts, and thus, the linguistic DIF detected here might be confined to a subset of the population. In Appendix D, we conduct analyses using bilinguals who reported English or Spanish as their first language, and find minimal differences in the extent to which DIF is detected. If English-first bilinguals are comparable to their English-dominant counterparts and Spanish-first bilinguals

[15]Given the lack of DIF, the visual items in the efficacy scale could potentially be used as "anchor items" to link English and Spanish scales. The use of visual items to enable linking translated forms is a promising avenue for future research.

are more similar to Spanish-dominant Latinos, these findings should allay concerns that our evidence of DIF only holds for certain kinds of Latinos. Even then, bilinguals comprise a large proportion of the Latino population, and thus, addressing DIF should yield scaling improvements for a sizable percentage of survey respondents.

It is also worth noting that concerns about representativeness are not unique to our method. Approaches to reducing translation errors such as back-translation also depend on a possibly unrepresentative set of expertly trained bilinguals. As Pérez (2009) notes, survey questions may be affected by the socioeconomic composition of translators who are proficient in both languages. Here we leverage a much larger set of bilinguals than those employed on translation teams, which might better reflect the diversity of the population. Moreover, an advantage of our design is that we can estimate counterfactual quantities via random assignment. Due to this study feature, we can estimate differences between items that consciously escape the notice of translation experts.

Our findings suggest that pre-testing items with bilinguals after translation has taken place may help identify items that suffer from DIF, which can enable scholars to construct scales that are more comparable across languages. Although DIF was prevalent across scales and items, we were unable to reject the null hypothesis of measurement equivalence in several cases. As Sireci and Berberoglu (2000) note, items that fail to display DIF in a bilingual sample may serve as good candidates for anchor items. DIF tests like those conducted in this study can also aid survey researchers and scholars in deciding which items to prioritize (i.e., those that do not suffer from DIF) when they face time or space constraints on surveys. By addressing DIF, scholars can be more confident in the inferences they draw from multilingual surveys.

References

Abrajano, Marisa. 2015. "Reexamining the "racial gap" in political knowledge." *The Journal of Politics* 77(1): 44–54.

Ankenmann, Robert D, Elizabeth A Witt, and Stephen B Dunbar. 1999. "An investigation of the power of the likelihood ratio goodness-of-fit statistic in detecting differential item functioning." *Journal of Educational Measurement* 36(4): 277–300.

Atar, Burcu, and Akihito Kamata. 2011. "Comparison of IRT Likelihood Ratio Test and Logistic Regression DIF Detection Procedures." *Hacettepe University Journal of Education* 41: 36–47.

Boroditsky, Lera. 2001. "Does language shape thought?: Mandarin and English speakers' conceptions of time." *Cognitive psychology* 43(1): 1–22.

Brislin, Richard W. 1970. "Back-translation for cross-cultural research." *Journal of cross-cultural psychology* 1(3): 185–216.

Ervin, Susan, and Robert T Bower. 1952. "Translation problems in international surveys." *Public Opinion Quarterly* 16(4): 595–604.

Finch, Holmes. 2005. "The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio." *Applied Psychological Measurement* 29(4): 278–295.

Flores, Alejandro, and Alexander Coppock. 2018. "Do bilinguals respond more favorably to candidate advertisements in English or in Spanish?" *Political Communication* 35(4): 612–633.

Hidalgo-Montesinos, M Dolores, and Juana Gómez-Benito. 2003. "Test Purification and the evaluation of differential item functioning with multinomial logistic regression." *European Journal of Psychological Assessment* 19(1): 1.

Hill, Kevin A, and Dario V Moreno. 2001. "Language as a variable: English, Spanish, ethnicity, and political opinion polling in South Florida." *Hispanic Journal of Behavioral Sciences* 23(2): 208–228.

Huddy, Leonie, Lilliana Mason, and Lene Aarøe. 2015. "Expressive partisanship: Campaign involvement, political emotion, and partisan identity." *American Political Science Review* 109(1): 1–17.

Iyengar, Shanto. 1993. "Assessing linguistic equivalence in multilingual surveys." *Social Research in Developing Countries: Surveys and Censuses in the Third World. London: John Wiley & Sons* , 173–182.

Kalla, Joshua L, and David E Broockman. 2020. "Reducing exclusionary attitudes through interpersonal conversation: evidence from three field experiments." *American Political Science Review* , 1–16.

King, Gary, and Jonathan Wand. 2007. "Comparing incomparable survey responses: Evaluating and selecting anchoring vignettes." *Political Analysis* 15(1): 46–66.

Lee, Taeku, and Efrén O Pérez. 2014. "The persistent connection between language-of-interview and Latino political opinion." *Political Behavior* 36(2): 401–425.

Lien, Pei-te, M. Margaret Conway, and Janelle Wong. 2003. "The contours and sources of ethnic identity choices among Asian Americans." *Social Science Quarterly* 84(2): 461–481.

Marian, Viorica, and Margarita Kaushanskaya. 2004. "Self-construal and emotion in bicultural bilinguals." *Journal of Memory and Language* 51(2): 190–201.

Nering, Michael L, and Remo Ostini. 2011. *Handbook of polytomous item response theory models*. Taylor & Francis.

Pérez, Efrén O. 2009. "Lost in translation? Item validity in bilingual political surveys." *The Journal of politics* 71(4): 1530–1548.

Pérez, Efrén O. 2011. “The origins and implications of language effects in multilingual surveys: A MIMIC approach with application to Latino political attitudes.” *Political Analysis* 19(4): 434–454.

Pérez, Efrén O, and Margit Tavits. 2016. “Language shapes public attitudes toward gender equality.”.

Pérez, Efrén O, and Margit Tavits. 2017. “Language Shapes People’s Time Perspective and Support for Future-Oriented Policies.” *American Journal of Political Science* .

Pérez, Efrén O, and Margit Tavits. 2019. “Language Influences Public Attitudes Toward Gender Equality.” *The Journal of Politics* 81(1): 81–93.

Pietryka, Matthew T., and Randall C. Macintosh. 2017. “ANES scales often don’t measure what you think they measure–An ERPC2016 analysis.”.

Prior, Markus. 2014. “Visual political knowledge: A different road to competence?” *The Journal of Politics* 76(1): 41–57.

Ramakrishnan, Karthick, and Farah Z Ahmad. 2014. “Language diversity and English proficiency.” *Center for American Progress* 27.

Sireci, Stephen G. 1997. “Problems and issues in linking assessments across languages.” *Educational Measurement: Issues and Practice* 16(1): 12–19.

Sireci, Stephen G, and Giray Berberoglu. 2000. “Using bilingual respondents to evaluate translated-adapted items.” *Applied measurement in education* 13(3): 229–248.

Slobin, Dan I. 1996. “From “thought and language” to “thinking for speaking”.”.

Welch, Susan, John Comer, and Michael Steinman. 1973. “Interviewing in a Mexican-American community: An investigation of some potential sources of response bias.” *The Public Opinion Quarterly* 37(1): 115–126.