

ARTICLES

Comparing Techniques to Identify Spanish-Speaking Households

Martha McRoy¹, Juanita Vivas Bastidas^{1,2}

¹ Methodology and Quantitative Social Sciences, NORC at the University of Chicago, ² Sociology, Loyola University Chicago

Keywords: Spanish, Bayesian Improved Surname Geocoding, BISG, survey language

<https://doi.org/10.29115/SP-2024-0022>

Survey Practice

Vol. 19 Special Issue, 2025

In the United States, recruiting representative samples of the Hispanic population has been challenging due to social factors such as culture, language, and trust levels. However, reaching this population is crucial for ensuring the representativeness and accuracy of public opinion data, particularly because this group has unique perspectives on topics including immigration, education, family, and healthcare. Previous research has explored methods for identifying Spanish-speaking households and improving panel recruitment strategies targeting Hispanic populations. This study compares three techniques used to identify Spanish-speaking households for a 2023 nationally representative survey of over 35,000 respondents. Specifically, we assess the accuracy of using vendor-provided data, Bayesian Improved Surname Geocoding (BISG), and Census Spanish language isolation locations in identifying likely Spanish-speaking households. Using sensitivity, specificity, and predictive values screening tests, we show that the BISG flag is the most sensitive technique, identifying the highest proportion of households where a respondent completed the survey in Spanish (82%). We conclude by providing recommendations for future surveys on identifying likely Spanish-speaking households.

Overview

Reaching Spanish-speaking households is crucial for ensuring the inclusivity and accuracy of public opinion data in the United States. These households have viewpoints and experiences distinct from non-Spanish-speaking populations and English-speaking Hispanics, particularly concerning immigration, education, and healthcare (Abrajano and Singh 2009; Escobedo, Cervantes, and Havranek 2023; Lee et al. 2012). Incorporating their perspectives contributes to a more comprehensive understanding of the attitudes, beliefs, and needs in America.

Researchers have explored methods to identify Spanish-speaking households and enhance panel recruitment strategies targeting Hispanic populations (Trussell 2010; Ventura et al. 2017). Trussell (2010) noted a strong preference among households where Spanish is predominantly spoken for bilingual recruitment materials. Ventura et al. (2017) recommended adjustments to ensure translations do not cause miscommunication, including reducing the quantity of text and simplifying the language used.

Studies have examined big data classifiers (BDCs) to predict the likelihood that a given household contains someone who is Hispanic (Dutwin et al. 2023; English et al. 2019; Suzuki et al. 2023). English et al. (2019) demonstrated that incorporating vendor data into the creation and upkeep of address-based sampling frames substantially improves the quality and coverage of address lists, leading to more accurate and representative survey samples. Suzuki et al. (2023) explored Bayesian Improved Surname Geocoding (BISG) as a cost-effective alternative to acquiring demographic data from vendors, demonstrating BISG’s potential to enhance the efficiency of surveys.

In this study, we examine various techniques to identify Spanish-speaking households for sending bilingual materials. Specifically, we evaluate the accuracy of vendor data, BISG flags, and Census Spanish linguistic isolation information¹ in identifying likely Spanish-speaking households. We compare the techniques using sensitivity, specificity, positive predictive values, and negative predictive values screening tests. Our results shed light on various methods to identify Spanish-speaking households prior to data collection.

Data and Analysis

Our data come from the 2023 Religious Landscape Study, a nationally representative address-based sample (ABS) mixed-mode survey of over 35,000 respondents from the Pew Research Center conducted from July 17, 2023, to March 4, 2024. Interviews were conducted via web, phone, and paper-and-pencil instrument with a 19.4% response rate². Fifteen percent of sampled households were identified as likely having at least one Spanish-speaker. These households were identified through a combination of methods:

1. Vendor Data: All households with at least one person flagged as likely Spanish speaking through available vendor data (vendor language data available for 95.4% of the sample),
2. BISG Predictions: All households predicted as having at least one member with a 50% chance or higher of being Hispanic determined through Bayesian Improved Surname Geocoding (BISG)³ when vendor language data were not available, and

1 Linguistic isolation is represented by “limited English-speaking households” in which no member 14 years old and over speaks only English or speaks both a non-English language and speaks English “very well” as defined by the American Community Survey (ACS) at the census tract level.

2 AAPOR Response Rate 1.

3 Bayesian Improved Surname Geocoding (BISG) was developed by RAND to produce accurate, cost-effective estimates of race and ethnicity. The approach efficiently combines two commonly used methods to estimate race and ethnicity: geocoded address and surname. <https://www.rand.org/health-care/tools-methods/bisg.html>

3. Spanish linguistically isolated Census Tracts: We calculated the proportion of Spanish-speaking households of the Census tract for each household. Households are sorted in descending order based on this proportion. Additional households are then flagged for Spanish language using this sort order after flagging households using vendor data and BISG predictions until 15% of the total sample has been flagged for bilingual mailings.

All sampled households received mailing materials in English. Spanish-flagged households received bilingual materials with the Spanish-language materials inserted in the envelopes so when opened, the Spanish-language materials would be seen first. Spanish households also received bilingual postcard reminders. All other households received English-only materials with a reference to Spanish materials available online and over the phone.

To measure the accuracy of these techniques to identify Spanish-speaking households, we calculated the sensitivity, specificity, and predictive values. [Figure 1](#) illustrates the derivations for these values.

Sensitivity measures the rate a technique identifies Spanish-speaking households from those that completed the survey in Spanish. It also calculates the technique's probability of not categorizing households as not having at least one Spanish-speaker when in fact they do have at least one.

Positive predictive value (PPV) calculates how many households flagged for Spanish completed the survey in Spanish. It also checks how many flagged households truly have a respondent who completed the survey in Spanish, avoiding incorrectly labeling non-Spanish-speaking households.

Specificity calculates how well a technique identifies households where a respondent did not complete the survey in Spanish. It assesses how many households are correctly identified, avoiding incorrectly labeling them as having a Spanish-speaking respondent.

Finally, negative predictive value (NPV) measures how likely it is that a household not flagged for Spanish completed the survey in English. It checks how many non-flagged households do not have a respondent who completed the survey in Spanish, avoiding incorrectly labeling them as having one.

We analyze our flags in the following order: vendor data Spanish language flag, BISG Hispanic flag, Census Spanish linguistic isolation flag, and the overall Spanish language flag (a flag generated by combining all three techniques). Notably, the vendor data flag analyses excluded cases in Connecticut, Colorado, and Virginia, as data were unavailable due to state regulations. By contrast, for the BISG analyses, we only included cases in these three states. Analyses were completed in Stata using sample weights that account for unequal probabilities of selection of addresses.

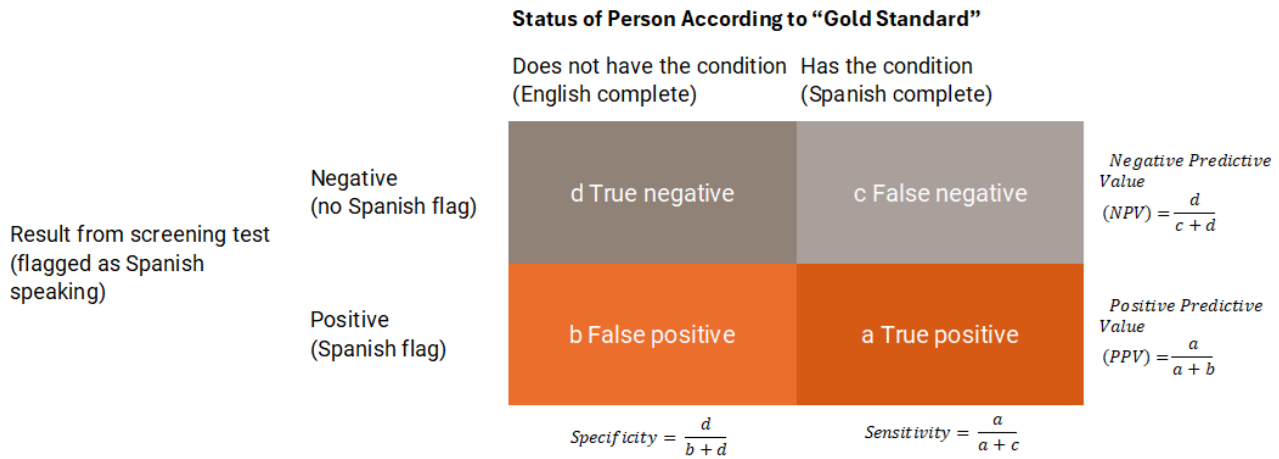


Figure 1. Diagram Demonstrating the Basis for Deriving Sensitivity, Specificity, and Positive and Negative Predictive Values. Adapted from Trevethan 2017.

Table 1. Weighted case counts and analytic tests for vendor data Spanish language flags.

Vendor Flag	Language of Interview		
	English	Spanish	
No Spanish Flag	31,552.3 (91.0%)	177.6 (0.5%)	Negative Predicted Value (NPV)= 99.4%
Yes, Spanish Flag	2,444.4 (7.1%)	495.0 (1.4%)	Positive Predicted Value (PPV)= 16.8%
	Specificity= 92.8%	Sensitivity= 73.6%	

Findings

Vendor Data Spanish Language Flags

Where vendor data were available, 11.2% of sampled households were flagged as likely having a Spanish speaker, resulting in 8.5% of responding households with the vendor flag for likely Spanish speakers. [Table 1](#) shows the accuracy measures for the vendor Spanish language flag. The flag's sensitivity value suggests that the technique correctly identified 73.6% of Spanish-speaking households among all households where a respondent completed the survey in Spanish, having a moderate ability to detect true positives (Spanish-speaking households). Its PPV indicates that, when the vendor data flag identified a household as likely having at least one Spanish-speaker, it was correct 16.8% of the time. According to the technique's specificity value, the vendor flag has a high probability of avoiding false positives as it correctly identified 92.8% of non-Spanish-speaking households among all households where there were no Spanish-speakers. Finally, its NPV reveals that, when the vendor flag identified a household as not likely having at least one Spanish-speaker, it was correct 99.4% of the time.

Table 2. Weighted case counts and analytic tests for BISG flags.

Bayesian Improved Surname Geocoding (BISG) Flag	Language of Interview		
	English	Spanish	
No Spanish Flag	1,991.8 (92.8%)	4.0 (0.2%)	Negative Predicted Value (NPV)= 99.8%
Yes, Spanish Flag	132.5 (6.2%)	18.1 (0.8%)	Positive Predicted Value (PPV)= 12.0%
	Specificity= 93.8%	Sensitivity= 82.0%	

Overall, the vendor flag has good specificity and NPV, indicating it performs well in correctly identifying non-Spanish-speaking households. The sensitivity is moderate (73.6%), suggesting it identifies a substantial proportion of true Spanish-speaking households but may miss some. The PPV is very low (16.8%), indicating that while it identifies Spanish-speaking households reasonably well, there are still some false positives.

Bayesian Improved Surname Geocoding (BISG) Hispanic Flags

The BISG Hispanic flag identified all households predicted as having at least one member with a 50% chance or higher of being Hispanic and was only run on states where vendor language data were unavailable. Given a strong association between ethnicity and household language, we flagged Hispanic households for Spanish materials. We flagged 9.5% of sampled households with the BISG flag, resulting in 7.0% of responding households in Connecticut, Colorado, and Virginia. [Table 2](#) displays the accuracy measures for this flag. The BISG flag's sensitivity value suggests that the technique correctly identified 82.0% of Spanish-speaking households among all households where a respondent completed the survey in Spanish, having a high probability of detecting true positives (Spanish-speaking households). Its PPV indicates that, when the BISG flag identified a household as likely having at least one Spanish-speaker, it was correct 12.0% of the time. The BISG flag's specificity value indicates it correctly identified 93.8% non-Spanish-speaking households among all households where there were no Spanish-speakers. Finally, its NPV reveals that, when the flag identified someone as a non-Spanish-speaker, it was correct 99.8% of the time.

The BISG Hispanic flag has a relatively high sensitivity, indicating it effectively identifies a large proportion of true Spanish-speaking households. The PPV is very low (12.0%), suggesting that when it identifies Spanish-speaking households, there is a notable rate of false positives. Finally, the flag demonstrates very high specificity and reasonably high NPV, indicating it performs well in correctly identifying non-Spanish-speaking households.

Table 3. Weighted case counts and analytic tests for Census Spanish Linguistic Isolation flags.

Census Flag	Language of Interview		
	English	Spanish	
No Spanish Flag	35,094.8 (95.3%)	459.5 (1.2%)	Negative Predicted Value (NPV)= 98.7%
Yes, Spanish Flag	1,026.2 (2.8%)	235.1 (0.6%)	Positive Predicted Value (PPV)= 18.6%
	Specificity= 97.2%	Sensitivity= 33.9%	

Census Spanish Linguistic Isolation Flags

The Census Spanish linguistic isolation information was used to flag sampled addresses from predominantly Spanish-speaking tracts for Spanish language materials. Households could be flagged with the Census flag and either the vendor flag or BISG flag. Overall, 5.0% of sampled households were flagged with the Census flag, resulting in 3.4% of responding households with this flag.

The accuracy measures for the Census Spanish linguistic isolation flag are found in [Table 3](#). The flag's sensitivity value suggests that the technique correctly identified 33.9% of Spanish-speaking households among all households where a respondent completed the survey in Spanish. Notably, the Census flag was the only one where there was a higher rate of Spanish interviews for households not flagged for Spanish than those that were flagged for Spanish (Sensitivity < 50%). Its PPV indicates that, when the flag identified a household as likely having at least one Spanish-speaker, it was correct 18.6% of the time. The flag's specificity value shows it correctly identified 97.2% of non-Spanish-speaking households among all households where there were no Spanish-speakers, meaning it has a high probability of avoiding false positives (non-Spanish-speaking households identified as Spanish-speaking households). Finally, its NPV reveals that the flag identified a household as not likely having at least one Spanish-speaker correctly 98.7% of the time.

The Census Spanish linguistic isolation flag has a low sensitivity (33.9%), indicating it identifies a reasonable proportion of true Spanish-speaking households. The PPV is relatively higher than the other flags (18.6%), suggesting that there is a higher probability of identifying true positives for Spanish-speaking households compared to other methods. Finally, the flag demonstrates high specificity and NPV, indicating it performs well in correctly identifying non-Spanish-speaking households.

Table 4. Weighted case counts and analytic tests for the combined overall Spanish flags.

	Language of Interview		
Overall Flag	English	Spanish	
No Spanish Flag	33,136.3 (90.0%)	118.6 (0.3%)	Negative Predicted Value (NPV)= 99.6%
Yes, Spanish Flag	2,984.7 (8.1%)	576.0 (1.6%)	Positive Predicted Value (PPV)= 16.2%
	Specificity= 91.7%	Sensitivity= 82.9%	

Overall Spanish Language Flags

Our study combined all three techniques to send Spanish language materials to 15% of sampled households. Overall, 9.7% of our flagged Spanish households completed the survey in Spanish, a higher rate than any of the individual techniques.

[Table 4](#) shows the accuracy measures for the full sample with all three techniques combined. The combined flag's sensitivity value suggests that by using all three techniques to flag Spanish-speaking households, we correctly identified 82.9% of Spanish-speaking households among all households where a respondent completed the survey in Spanish, having a high probability of detecting true positives. Its PPV indicates that, when the flag identified a household as likely having at least one Spanish-speaker, it was correct 16.2% of the time. Its specificity value indicates it correctly identified 91.7% of non-Spanish-speaking households among all households where there were no Spanish-speakers. Its NPV reveals that, when the overall flag identified a household as not likely having at least one Spanish-speaker, it was correct 99.6% of the time.

The overall Spanish language flag has a relatively high sensitivity, indicating it effectively identifies a large proportion of Spanish-speaking households. The PPV is very low (16.2%), suggesting that while it identifies Spanish-speaking households well, there are still a considerable number of false positives. Finally, we found the overall Spanish language flag to have good specificity and negative predictive values, indicating it also performs well in correctly identifying non-Spanish speaking households.

Conclusions

Comparing the accuracy of all techniques, we observed that the BISG Hispanic flag is the most sensitive (see [Table 5](#)). It identified the highest proportion of households where a respondent completed the survey in Spanish (82.0%). The Census Spanish linguistic isolation flag had the lowest sensitivity, identifying only 33.9% of true positives. Despite its low sensitivity, the Census flag had the highest PPV (18.6%), indicating it is better at ensuring that positive results are true positives. The Census flag also had the

Table 5. Sensitivity, Specificity, and Predictive Values by technique.

	Overall flag	Vendor flag	Bayesian Improved Surname Geocoding (BISG) flag	Census flag
Sensitivity	82.9%	73.6%	82.0%	33.9%
Positive Predicted Value (PPV)	16.2%	16.8%	12.0%	18.6%
Specificity	91.7%	92.8%	93.8%	97.2%
Negative Predicted Value (NPV)	99.6%	99.4%	99.8%	98.7%

highest specificity (97.2%), indicating it is the best at correctly identifying negatives. Finally, the BISG flag had the highest NPV (99.8%), suggesting it is the most reliable technique at identifying true negatives.

For our purposes, it was better to have higher rates of false positives than higher rates of false negatives since we would rather send bilingual materials to households where there are no Spanish-speakers than send English-only materials to households where people prefer completing the survey in Spanish. Nonetheless, respondents receiving English-only materials were still able to complete the survey in Spanish online or on the phone.

With that in mind, we conclude that the best individual technique in identifying Spanish-speaking households is the BISG Hispanic flag as this method is the most sensitive and has the highest NPV, making it a strong choice for identifying as many true positives as possible and correctly identifying true negatives. However, it has the lowest PPV, meaning it might have more false positives. Using the overall Spanish language flag by combining the three flags has the highest sensitivity (82.9%) and second highest NPV (99.6%), showcasing that it can be best to use multiple techniques to identify likely Spanish-speaking households.

There may be within-household heterogeneity in race and ethnicity – while someone in the household may be Hispanic, others may self-classify as non-Hispanic. It may also be the case that someone in the household can complete the instrument in English but may not speak only English at home.

Recommendations

Given that for our purpose it is better to have higher rates of false positives than higher rates of false negatives, we identified that the BISG Hispanic flag is the best single technique for identifying Spanish-speaking households.

We only used the BISG Hispanic flag in states where vendor language data were not available. As such, future studies could use the BISG Hispanic flag to identify households likely having at least one Spanish-speaker in all states. Importantly, this option still entails purchasing limited vendor data

for surnames for the algorithm as the BISG flag is generated from the combination of two methods to estimate race and ethnicity: geocoded address and surname.

When using vendor data, we examined language used within household, but race and ethnicity information is also available. Future research on using vendor data race and ethnicity to identify Spanish-speaking households should be explored.

Where there is a budget or time constraint, we recommend using the Census linguistic isolation flag. This method is free and quick, as it only requires researchers to merge ACS data to geocoded addresses. It performed well in ensuring that positive results are true positives and correctly identifying non-Spanish-speaking households. It also works for different languages.

Future research can apply these techniques for targeting hard-to-survey populations or other languages as vendor and Census data are available for more subgroups. The current BISG algorithm predicts six race and ethnicity categories, which may limit its utility for specific populations.

.....

Lead Author Contact Information

Martha McRoy
mcroy-martha@norc.org

Submitted: September 23, 2024 EST. Accepted: November 02, 2024 EST. Published: March 02, 2025 EST.

REFERENCES

- Abrajano, M., and S. Singh. 2009. "Examining the Link Between Issue Attitudes and News Source: The Case of Latinos and Immigration Reform." *Political Behavior* 31 (1): 1–30. <https://doi.org/10.1007/s11109-008-9067-8>.
- Dutwin, D., P. Coyle, J. Lerner, I. Bilgen, and N. English. 2023. "Leveraging Predictive Modelling from Multiple Sources of Big Data to Improve Sample Efficiency and Reduce Survey Nonresponse Error." *Journal of Survey Statistics and Methodology* 12 (2): 435–57. <https://doi.org/10.1093/jssam/smad016>.
- English, N., T. Kennel, T. Buskirk, and R. Harter. 2019. "The Construction, Maintenance, and Enhancement of Address-Based Sampling Frames." *Journal of Survey Statistics and Methodology* 7 (1): 66–92. <https://doi.org/10.1093/jssam/smy003>.
- Escobedo, L. E., L. Cervantes, and E. Havranek. 2023. "Barriers in Healthcare for Latinx Patients with Limited English Proficiency—a Narrative Review." *Journal of General Internal Medicine* 38 (5): 1264–71. <https://doi.org/10.1007/s11606-022-07995-3>.
- Lee, S. M., A. Thorn, S. C. Bloomdahl, J. H. Ha, S. K. Nam, and J. Lee. 2012. "Parent Involvement in School: English Speaking versus Spanish Speaking Families." *The Spanish Journal of Psychology* 15 (2): 582–91. https://doi.org/10.5209/rev_SJOP.2012.v15.n2.38869.
- Suzuki, H., M. McRoy, N. Connor, and J. Brinkley. 2023. "Bayesian Improved Surname Geocoding as an Efficient Alternative to Purchased Race and Ethnicity Data for Health Surveys." <https://doi.org/10.21203/rs.3.rs-2887048/v1>.
- Trevethan, R. 2017. "Sensitivity, Specificity, and Predictive Values: Foundations, Plabilities, and Pitfalls in Research and Practice." *Frontiers in Public Health* 5:307. <https://doi.org/10.3389/fpubh.2017.00307>.
- Trussell, N. 2010. "Spanish Respondents' Choice of Language: Bilingual or English?" *Survey Practice* 3:1–4. <https://doi.org/10.29115/SP-2010-0003>.
- Ventura, I., R. Bautista, D. Gleicher, C. Milesi, and E. Hendarwan. 2017. "Panel Recruitment for Spanish Speaking Populations: The AmeriSpeak® Case Study." In *Proceedings of the American Statistical Association, Survey Research Methods Section*, 3754–60. <http://www.asasrms.org/Proceedings/y2017/files/594088.pdf>.