

Replication of “Race/Ethnicity and Family Contact: Toward a Behavioral Measure of Familialism”¹

Anna Rakes

Elon University

STS 3300: Statistical Methods for Data Analytics

Professor Nicholas Bussberg

2024-01-22

¹ Author can be contacted at arakes@elon.edu. The author would like to thank Professor Nicholas Bussberg for their guidance and assistance in the paper.

ABSTRACT

Joseph Comeau in his article “Race/Ethnicity and Family Contact: Toward a Behavioral Measure of Familialism”, aims to investigate the effect of ethnicities on familialism, and more specifically, add to prior research that suggests Latinos in the United States share more familialism within their families than other ethnicities. Existing research used attitudinal measures to conclude that Latinos have higher levels of familialism, whereas Comeau uses data from the 2002 General Social Survey to measure familialism through the frequency of contact with family members. Comeau’s behavioral measure of familialism revealed that Latinos have more contact with their family than White non-Hispanics. My goal was to use Comeau’s methods found in his article and the GSS data to replicate the same results as Comeau without knowledge of his exact data cleaning and analysis process. While I achieved similar overall descriptive statistics and mostly similar regression model results, due to the presence of multiple race variables in the GSS survey and multiple ways to define the race of the respondents, I did not find that Hispanics have more contact with their nuclear family than White non-Hispanics do. However, I had similar results for individual family members as well as for the extended family index.

INTRODUCTION

Past literature suggests that familialism, which is defined by Comeau as “a sense of loyalty, identification, solidarity, and attachment to both nuclear and extended family”, plays a vital role in Latino culture in America (Comeau, 2012). More specifically, familialism is seen to be stronger after first-generation Latino families in the US due to a small family presence in first-generation families. Most research surrounding this topic has concluded that since familialism is a crucial component of Hispanic culture, attitudinal and behavioral measures of familialism are

higher in Hispanics than other races. There is, however, conflicting research around whether Black non-Hispanics have similar levels of familialism as Hispanics. Existing research on familialism shows that attitudinal measures of familialism are more accurate than behavioral measures, but Comeau looks to Keefe's research of behavioral measures of familialism in Hispanics to improve his own and add to the behavioral literature. Unlike Keefe's work, which only looks into Mexican Americans and Anglo Americans' frequency of visits to family households, Comeau controls for factors such as age, sex, education, and geographic mobility, and investigates overall family levels of familialism as well as individual family member categories, such as mother, father, cousins, sibling, etc, to effectively reveal insights between ethnicity and behavioral familialism levels. To analyze behavioral measures, Comeau looks at the frequency of contact between the respondent and their nuclear family, extended family, as well as individual family members.

For Comeau's research, he uses data from the 2002 General Social Survey. Since his interest is in comparing levels of familialism among Hispanics, White non-Hispanics, and Black non-Hispanics, those are the only 3 races that he gathers his data from. Comeau's response variables in his analysis are the family categories, which consist of two overall family indexes (nuclear and extended) and 7 individual family categories (mother, father, sibling, adult child, cousin, uncles/aunts, and nieces/nephews). For the individual nuclear family categories, consisting of mother, father, sibling, and adult child, the responses ranged from a scale of 1 to 6, where 1 indicates the respondent visits the family members less often/never and 6 indicates the respondent visits the family member daily. The overall nuclear family index variable used in the analysis is an average of the frequency of contact among the mother, father, sibling, and adult child variables. For the individual extended family categories, consisting of cousins,

uncles/aunts, and nieces/nephews, the responses ranged from a scale of 1 to 3, where 1 indicates the respondent hasn't visited the family member at all in the last 4 weeks, and 3 indicates the respondent visited the family member more than twice in the last 4 weeks. Similar to the nuclear family index, the extended family index is an average of the frequency of contact among cousins, uncles/aunts, and nieces/nephews.

The explanatory variables used in Comeau's analysis include race, gender, marital status, geographic mobility, education, age, and income. Comeau turned race into two dummy variables of White non-Hispanic and Black non-Hispanic, with Hispanic as the reference category for both. Gender was also a dummy variable with males as the reference category. Comeau also made marital status into a dummy variable, where responses for "never married" were the reference category, and any response indicated to having been married, such as "married", "widowed", "divorced", or "separated", was coded as 1 for the marital status variable. Geographical mobility was also changed by Comeau into two dummy variables. If the respondent still lives in the same city and state as they did when they were 16, it was treated as the reference category for both of the geographic mobility variables. If the respondent lives in the same state but just a different city as they did when they were 16, it was coded as a 1 for the first geographic mobility variable. However, if the respondent lives in an entirely different state than they did when they were 16, the second variable is coded as a 1. Comeau only included data from those over the age of 18, because most of those under 18 are still living with their family and therefore questions of family contact are not applicable. Comeau left the education and income variables untouched. Overall, Comeau was left with a sample size of 766 for the nuclear family index and 771 for the extended family index.

Regarding Comeau's data analysis, he produced two tables, one of which contained descriptive statistics for each of the variables. Comeau's Table 1 contains the mean and standard deviation for the frequency of contact with each family member and the two family indexes. Each of the 7 family members and the 2 family indexes are treated as their own sample in Table 1. Within each sample, Comeau also computed the percentage of that sample that is white, Hispanic, and black. Similarly, there are percentages of respondents that are male, female, have ever been married, and from a different state or city than they were at 16, for each sample/variable. Within each sample/variable, the mean and standard deviation of age, education, and income are recorded.

Table 2 in Comeau's article produced numerous multiple linear regression models for the 7 individual family categories and the 2 family indexes. For each of the 9 total models, the predicted y-value is the frequency of contact with the corresponding family member and the regression coefficients are given as "*b*" for each of the following variables: White non-Hispanic, Black non-Hispanic, different city from age 16, different state from age 16, sex, age, ever been married, education, and income. Comeau also provides the standard error for each of the variables in each model as well as the p-value. The p-values for the race variable were interpreted as a one-tailed test, denoted by a [†] for a p-value less than 0.05, ^{††} for p-values less than 0.01, and ^{†††} for p-values less than 0.001. The remaining variables were interpreted as two-tailed tests, denoted by a * for a p-value less than 0.05, ** for a p-value less than 0.01, and *** for a p-value less than 0.001.

In Comeau's analysis, he found that Hispanics have more frequent contact with both nuclear and extended family members than White non-Hispanics. However, in terms of frequency of contact among Hispanics and Blacks, the only family member that Hispanics had

more contact with than Blacks was their father. This means that overall, Hispanics and Black non-Hispanics have similar levels of family contact and familialism. In terms of each individual family category, all the models were found to be significant, with the adult child model being the weakest. Regarding the variables other than race, being in a different state from the age of 16 had a significant negative effect on each of the 9 models, suggesting those who move to a different state than where they lived when they were 16 are less likely to contact both their nuclear and extended family members.

DATA AND METHODS

To begin the process of replicating Comeau's results, I loaded the GSS dataset into R Studio. I narrowed down the large GSS dataset to just the observations I wanted by filtering the year variable to only 2002 and the age variable to only 18 and older {arakes-repl-data1.R AMR 2024-01-14}. Additionally, I subset the data to only include the variables used in Comeau's analysis, described in detail on pages 255-257 of the original article {arakes-repl-data1.R AMR 2024-01-14}. These variables included the following: year, wtssall, RACECEN1, RACECEN2, RACECEN3, age, sibvisit, kidvisit, PAVISIT1, MAVISIT1, cousins, uncaunts, niecenep, sex, MOBILE16, educ, and RINCOM98. In the process of determining which race variable from the GSS survey I wanted to use, I discovered multiple variables that could have been used by Comeau to subset his data by race. For example, the GSS has a simple variable race, but the only options for that variable were "white", "black", or "other". I felt as though this would not be able to be used for my analyses since the focus is on Hispanics. Additionally, the GSS specified that after the year 2000, the variables "RACECEN1", "RACECEN2", and "RACECEN3" were used to determine race. Since Comeau's analysis was from 2002, I figured these variables would be the best fit. In addition, the RACECEN variables included a specific answer for Hispanics,

unlike the original race variable. Once I decided on using the RACECEN variables, I subset my data to only include responses to the RACECEN1 variable that were White, Black, or Hispanic {arakes-repl-data1.R AMR 2024-01-14} . Comeau's analyses also referred to White and Black respondents as non-Hispanic, so I subset the RACECEN2 and RACECEN3 variables to not include Hispanics {arakes-repl-data1.R AMR 2024-01-14}. These variables were available for people who identified with more than one race. This way, I would only have respondents in my dataset that were White, Black, or Hispanic, and no one that was White and Hispanic or Black and Hispanic.

Once I had narrowed down the GSS dataset to only include observations used in Comeau's analysis, I was able to create the dummy variables. I discovered some confusion within Comeau's description of the race dummy variables because he referred to them on page 256 as two dummy variables for White and Black with Hispanic as the reference category for each. However, in Table 1 on page 259, he calculated his samples as a percentage of all three races, suggesting there was either a third dummy variable where Hispanic was coded as 1 or a third variable that was not a dummy and included all 3 races. To combat this, I still created two dummy variables where the first one had White non-Hispanics coded as 1 and Hispanics coded as 0 and the second one had Black non-Hispanics coded as 1 and Hispanics coded as 0, however, I created a third variable to be used in my replication of Table 1 that was not a dummy variable. This variable had Hispanics coded as 0, White non-Hispanics coded as 1, and Black non-Hispanics coded as 2 {arakes-repl-data2.R AMR 2024-01-16}. To mimic Comeau's dummy variable for sex, I coded females as 1 and males as the reference category of 0 {arakes-repl-data2.R AMR 2024-01-16}. This was flipped from the GSS data, where females had taken on a value of 2 and males a value of 1. Similarly, to mimic Comeau's dummy variable for marital

status, I coded “married”, “widowed”, “divorced”, and “separated” all as 1 to represent if the respondent had ever been married. “Never married” was coded as 0 for the reference category {arakes-repl-data2.R AMR 2024-01-16}. The last dummy variables I created represented geographic mobility. Comeau said he created two dummy variables, each of which, if the respondent lived in the same city as they did when they were 16, it was treated as the reference category and coded as 0. I created the first dummy variable titled “diffcityB” which said that if the respondent lived in same state but different city than they did when they were 16, it would be recoded as a 1 {arakes-repl-data2.R AMR 2024-01-16}. Similarly, the second dummy variable, titled “diffstateB” said that if the respondent lived in an entirely different state than they did when they were 16, it would be recoded as 1 {arakes-repl-data2.R AMR 2024-01-16}. Both “diffcityB” and “diffstateB” received 0 if the respondent still lived in the same state and same city as they did when they were 16.

The next step in cleaning my dataset was to recode each of the family category variables. As mentioned previously, Comeau flipped the scale for each of the family variables as well as combining a few of the responses and assigning a few as NA. For each of nuclear family variables (kidvisit, sibvisit, PAVISIT1, and MAVISIT1) I created a new variable that assigned values of 1 to a 6 and values of 2 to a 6, since Comeau combined responses of “she or he lives in the same household as I do” which was previously a 1, and “daily” which was previously a 2, into one response coded as 6 {arakes-repl-data3.R AMR 2024-01-16}. I followed Comeau’s instructions on pages 255 and 256 to recode the rest of the values for the nuclear and extended family variables. The last step in cleaning the data was to create the nuclear and extended family indexes, which are averages of the individual family variables. The variable “nucfamily” took an average of the responses from the new “kidvisit2”, “sibvisit2”, “pavisit2”, and “mavisit2”

variables {arakes-repl-data4.R AMR 2024-01-16}. NA values were dropped from the average of each row. The extended family index was also created into a new variable called “extfamily” which took an average of the responses from the new “cousins2”, “uncaunts2”, and “niecenep2” variables and dropped the NA values {arakes-repl-data5.R AMR 2024-01-17}. The variables from the original GSS dataset were also coded as character variables, so I had to convert the variables that would be used in the regression model, age, educ, and RINCOM98, into numerical variables.

To confirm that my data was clean and selected correctly, I ran diagnostics to compare sample sizes for each of the variables {arakes-repl-data5.R AMR 2024-01-17}. On page 259 in Table 1, Comeau reported a sample size of 766 for the nuclear family index and a sample size of 771 for the extended family index. In comparison, my final dataset reported a sample size of 1,070 for the nuclear family index and 1,082 for the extended family index, as seen in Table A, my replicated table of descriptive statistics {arakes-repl-data5.R AMR 2024-01-17}. The slight discrepancy in sample size could be due to the difference in filtering the dataset based on race. Since there were many possible ways to calculate and subset the multiple race-related variables, the results will be slightly off. My use of RACECEN1 to subset race led to an overrepresented proportion of white respondents and an underrepresented proportion of Hispanic respondents. While Comeau’s Table 1 reveals that White non-Hispanics make up 77.6% of the nuclear family index, my replicated Table A reports that White non-Hispanics make up 83.8% {arakes-repl-analysis1.R AMR 2024-01-18}. This pattern is seen through the remaining individual family variables.

Despite the difference in the racial makeup of the sample, the overall mean and standard deviations for each of the family variables are within 0.1. For example, Comeau reported in

Table 1 a mean contact frequency for the nuclear family index of 3.468, whereas I reported in Table A a mean of 3.42 {arakes-repl-analysis1.R AMR 2024-01-18}. Additionally, Comeau reported in Table 1 a mean contact frequency for the extended family index of 1.843, whereas I reported in Table A a mean of 1.773 {arakes-repl-analysis2.R AMR 2024-01-18}.

The only other significant difference is the proportion of females/males that make up the samples. On average, Comeau's descriptive statistics reveal that males make up about 52% of his samples. Whereas my descriptive statistics reveal that females make up about 51% of my samples {arakes-repl-analysis1.R AMR 2024-01-18}. The differences between the gender make up could be due to the general size difference between the samples. Since the sex variable is very close to an even proportion of females and males in Comeau's sample, it is possible for fluctuation among the makeup of respondents' sex, as seen in my replicated Table A. Overall, the means, standard deviations, and percentages in Table A are all within 0.1 of Table 1, except for the race and sex variables.

RESULTS AND DISCUSSION

Once I had cleaned my data and verified that my sample is similar to that of Comeau's, I was able to replicate Table 2 on page 261, which produced 9 multiple linear regression models for each of the 7 individual family variables and the 2 family indexes. Each model predicted the associated family category frequency based on the two race dummy variables, the two geographic mobility dummy variables, sex, age, marital status, education, and income. Within each model on page 261, Comeau reported the regression coefficient, standard error, p-value, R^2 value, F-test statistic, and the sample size. Table B represents my replicated table of regression models. Overall, Comeau found that the strongest model was the mother model with an R^2 value of .300 and the weakest was the nephew/niece model with an R^2 value of .070. Similarly, my

strongest model in Table B was the mother model with an R^2 value of 0.269 and the weakest was the nephew/niece model with an R^2 value of 0.062 {arakes-repl-analysis4.R AMR 2024-01-18}.

Comeau found that for the nuclear family index model, the following variables were significant in the model at the .05 level: White non-Hispanics, living in a different city from age 16, living in a different state from age 16, marital status, and the f-test statistic. My replicated Table B found that the following variables were significant in the model at least at the 0.05 level: Black non-Hispanics, living in a different city from age 16, living in a different state from age 16, marital status, and the f-test statistic {arakes-repl-analysis4.R AMR 2024-01-18}. In comparison, the only difference in significant variables between Comeau's model and mine was the race variable. Comeau's model concludes that Latinos have a higher frequency of contact with extended family members than White respondents, however, my model concludes that Latinos have a higher frequency of contact with extended family members than Black respondents. Once again, this difference in results related to the race variable can be attributed to the different methods taken to create the race variable.

For the extended family index model on page 261, Comeau found the following variables significant at least at the 0.05 level: White non-Hispanics, living in a different city from age 16, living in a different state from age 16, age, and the f-test statistic. My extended family index model in Table B found the same variables significant at the 0.05 level as well {arakes-repl-analysis4.R AMR 2024-01-18}.

In terms of the individual family models, the mother, sibling, and cousin models on page 261 and my mother, sibling, and cousin model in Table B each had the same variables significant at the 0.05 level. However, within the father, adult child, aunt/uncle, and nephew/niece models there were a few discrepancies between Comeau's models and mine in Table B. These

discrepancies were only found within the significance or lack thereof the race and sex variables. I believe the difference in the significance of the race variables for my father, adult child, and aunt/uncle model is due to the difference between Comeau's and my creation of the race variable. The differences between the significance of the sex variable for the father and adult child model could, once again, be due to the general size difference between the samples and the close gender makeup of the sample. Despite the few differences in the significance of the race and sex variables within my replicated regression models, the overall significance of our models match. In general, the coefficients and standard error for each of the models on page 261 are within tenths of my replicated models in Table B{arakes-repl-analysis4.R AMR 2024-01-18}. In addition, the majority of the variables that Comeau found to be significant in each of his models were also found to be significant in each of my models.

CONCLUSION

Comeau concluded through his analyses that Latinos have a higher frequency of contact with both nuclear family and extended family than White non-Hispanics have. However, there was no statistically significant difference between the frequency of contact for Hispanics and Black non-Hispanics. This finding is consistent with that of Hays and Mindel's work, which revealed that White non-Hispanics have less contact with their family than Black non-Hispanics and Hispanics have. Unfortunately, due to the multitude of paths that could have been taken to gather a sample of Hispanics, White non-Hispanics, and Black non-Hispanics, I had an overrepresented proportion of White non-Hispanics in my sample. This difference in racial makeup led to a few of my models reaching different conclusions. While Comeau concluded that Latinos have more contact with nuclear family members than White non-Hispanics, I found that Latinos actually have more contact with nuclear family members than Black non-Hispanics. I

did, however, come to a similar conclusion as Comeau did in terms of extended family. Both of our models reveal that Latinos have more contact with extended family members than White non-Hispanics have. Within individual family categories, Comeau reported that Latinos have more contact with fathers, adult children, aunts/uncles, nephews/niece, and cousins than White non-Hispanics have. My regression models revealed that Latinos have more contact with aunts/uncles, nephews/niece, and cousins than White non-Hispanics have.

Despite the differences in conclusions for the frequency of contact in the nuclear family, father, and adult child model, I was able to replicate similar results to Comeau. Our descriptive statistics revealed a similar makeup of the sample, except for race and sex. Our regression models revealed similar effects and significance of geographic mobility, age, marital status, education, and income on the frequency of contact with family members as well as similar overall strength and significance of each of the models.

BIBLIOGRAPHY

Comeau, J. A. (2012). Race/Ethnicity and Family Contact: Toward a Behavioral Measure of Familialism. *Hispanic Journal of Behavioral Sciences*, 34(2), 251-268.
<https://doi.org/10.1177/0739986311435899>