Bates, Nancy, and Joe Zamadics. 2021. "COVID-19 Infection Rates and Propensity to Self-Respond in the 2020 U.S. Decennial Census." *Survey Practice* 14 (1). https://doi.org/10.29115/SP-2021-0002.

ARTICLES

COVID-19 Infection Rates and Propensity to Self-Respond in the 2020 U.S. Decennial Census

Nancy Bates[1] [a], Joe Zamadics[2]

[1] US Census Bureau, [2] PSBinsights

Keywords: covid-19 infection rates, self-response rates, decennial census

https://doi.org/10.29115/SP-2021-0002

Survey Practice

Vol. 14, Issue 1, 2021

Survey methodologists have acknowledged that the social environment may influence survey and census participation—both at the societal level and at the community level. The 2020 COVID-19 pandemic affected daily life throughout the entire U.S., but to differing degrees depending upon the particular neighborhood or community. In this article, we use U.S. Centers for Disease Control county-level COVID-19 infection data coupled with U.S. 2020 Decennial Census response rate data to explore whether this societal level pandemic influenced participation in the census. We found that even when controlling for covariates predictive of COVID-19 infection rates (e.g., percent minority population, age 65+), infection rates were found to be significantly (negatively) associated with self-response.

Introduction

Survey methodologists have acknowledged that the social environment can influence survey and census participation—both at the societal level and at the community level (Groves and Couper 1998; Johnson et al. 2006). The 2020 COVID-19 pandemic affected daily life throughout the entire U.S., but to differing degrees depending upon neighborhood or community (Orgera, McDermott, and Rae 2020; University of Minnesota 2020). Mandates to quarantine and wear face coverings in public varied across states and among localities within states (e.g., cities vs. more rural places, see Bunks and Rough 2020; Weise 2020). In addition, population density, public transportation usage, and phased reopenings of local economies varied state to state and locale to locale (COVID-Local 2020; Lee et al. 2020; Moore and Lazar 2020; National Public Radio 2020; U.S. Chamber of Commerce 2020). Such factors are hypothesized to correlate with COVID-19 infection rates (Liu et al. 2020).

In early August, the Census Bureau announced it would conclude nonresponse follow-up (NRFU) operations by the end of September. This was a reversal from earlier plans to extend NRFU through the end of October (to make up for operational postponements due to the pandemic). As a result, the agency looked for innovative ways to complete the census count on the accelerated schedule. The goal of this article is to quantify whether a

a All of the data used in this study were taken from publicly available sources; therefore, the paper is not subject to a Census Bureau disclosure review. The views expressed in this article are those of the authors, and not necessarily the U.S. Census Bureau.

community-level environmental variable (operationalized as COVID-19 infection rates at the country level) was associated with participation in the U.S. 2020 Decennial Census.

Methodology

In mid-March, the Census Bureau mailed materials to households with instructions to complete the census online, by phone, or by paper questionnaire. Soon after, the agency announced a public facing tract-level response rate map to track (in close to real time) self-response to the 2020 Decennial Census (U.S. Census Bureau 2020). Response rate data for this article were extracted as of July 25, 2020. At that point, the cumulative national response rate was 64.6% with a standard deviation of 12.7 percentage points. Cumulative tract level response rates were merged with other tract level data from the Census Bureau's Planning Database (PDB)—a data file containing a subset of 2014–2018 5-year American Community Survey (ACS) estimates as well as other census operational variables. The PDB variables included those documented to predict 2010 Census self-response (Erdman and Bates 2017) as well as those correlated with COVID-19 death rates (Knittel and Ozaltun 2020). These included socioeconomic, household, and population density variables (see Appendix for entire list).

Finally, we included a variable indicating the mail implementation strategy each census tract was flagged to receive—either mailing flights that first encouraged online response without a paper questionnaire (Internet First) or flights that included a paper questionnaire in the first mailing (Internet Choice). Census tracts received a paper questionnaire in the first mailing (Internet Choice) if the area was expected to have lower Internet usage and thus would be more likely to benefit from an earlier paper questionnaire. Tracts were assigned to Internet Choice if they had lower self-response rates to the ACS and had either low Internet response, a higher population of people aged 65 or more, or low Internet subscribership. Otherwise, tracts were assigned to Internet First. The First/Choice tract-level indicator variable was available at the public facing website. The cumulative self-response rates, PDB variables, and First/Choice variable were then merged by tract.

Next, we merged these data with the latest COVID-19 cumulative infection rate data from the CDC. Rates were defined as the total number of positive COVID-19 tests since January in a given county over that county's total population. Thirty-nine counties contained no positive cases; the mean positive infection rate as of July 25, 2020 was 1.5%; and the maximum was McKinley County, NM, with a positive infection rate of 3.6%. The most granular COVID-19 data available was at the county level. Consequently, we paired tracts with their cumulative county infection rate as of July 25, 2020. Our analysis was limited to census tracts where all housing units within the tract were designated to receive 2020 Census materials in the mail by mid-

Table 1: COVID-19 county-level infection rate quartiles by 2020 Census self-response rates.*

Infection Quartile	Response Rate	Mean County-Level Infection Rate	Total Population
Lowest Quartile County Infection Rate	68.70%	0.53%	63,952,451
2nd Lowest Quartile County Infection Rate	68.20%	1.08%	64,844,088
2nd Highest Quartile County Infection Rate	62.70%	1.66%	67,795,789
Highest Quartile County Infection Rate	60.60%	2.91%	64,226,785

* Cumulative infection and response rates as of July 25, 2020

March.[1] Each model used ordinary least squares (OLS) to predict tract-level self-response rate as of July 25, 2020. In addition to the PDB predictors previously listed, a state fixed effect was included (output not shown).

Results

Table 1 shows response rates broken down by COVID-19 infection rate quartiles and demonstrates that response rates are lower in counties with high COVID-19 infection rates. The county infection rate column shows the average infection rate among counties in each quartile. For example, the average county in the lowest infection quartile had 0.53% of its population test positive for COVID-19 since the start of the pandemic, and on average, 68.7% of households responded to the census. From this bivariate perspective, response and infection rates appear to be negatively correlated, with the lowest infection rate quartile having the highest response rate (68.7%) and the highest infection rate quartile having the lowest response rate (60.6%).

Table 2 presents an OLS regression model predicting cumulative self-response (mail, online, and telephone response combined[2]). Results indicate that COVID-19 infection county rates were a significant (and negative) predictor of self-response, even when controlling for a variety of operational, socio-economic, and demographic covariates known to be associated with census participation and COVID-19 infection rates. The R^2 indicated that a significant portion of the variance of response rate was accounted for (at around 80%).

The cumulative response rate model shows a negative relationship between county-level COVID-19 cases and response. For every percentage point increase in a tract's county-level infection rate, the model expects response to fall about 1.3 percentage points. The standard deviation in county infection rate was 1.1 percentage points. Figure 1 shows this relationship visually. Most tracts were in counties with a cumulative infection rate below 5%, meaning the

1 78.8 percent of Census tracts were designated as this type of enumeration area. Examples of areas not designated to receive mailed materials include those that either do not have mail delivered to the physical location of the housing units and extremely remote areas such as southeast Alaska and select American Indian reservations that requested personal visit enumeration.

2 As of July 25, 2020, 62.5% of households responded to the census. Of the responses, 80% were completed on the Internet, 19% were paper responses, and 1% were completed by telephone.

Table 2: OLS regressions of 2020 Census Self-Response Rates (census tract level) on specified predictor variables

	Dependent variable
	Cumu. RR
Cases Rate	-1.332***
	-0.052
% Hispanic Alone	-0.023***
	-0.003
% Black African American in Combo	-0.044***
	-0.002
% Asian in Combo	-0.021***
	-0.003
% American Indian/ Alaskin Native in Combo	-0.207***
	-0.011
% White Low Education	-0.072***
	-0.002
% Renter Occupied	-0.179***
	-0.002
% Spanish Speakers	-0.136***
	-0.006
% Vacant Units	-0.728***
	-0.004
Median HH Income	0.00004***
	< 0.001
% Internet First	1.917***
	-0.075
% Age 65 +	0.166***
	-0.004
% Age 18-24	-0.105***
	-0.004
% Female House	0.014***
	-0.004
Density	0.016***
	(.002)
Observations	57,995
R2	0.79
Adjusted R2	0.79
Residual Std. Error	5.789 (df = 57,929)
F Statistic	3,347.745*** (df = 65; 57,929)

Note: *p < 0.05; **p < 0.01; ***p < 0.001.

predicted effect for most counties was under 3 percentage points. However, there were a significant number of tracts in areas such as New York City and parts of Arizona where the predicted effect was over 5 percentage points.

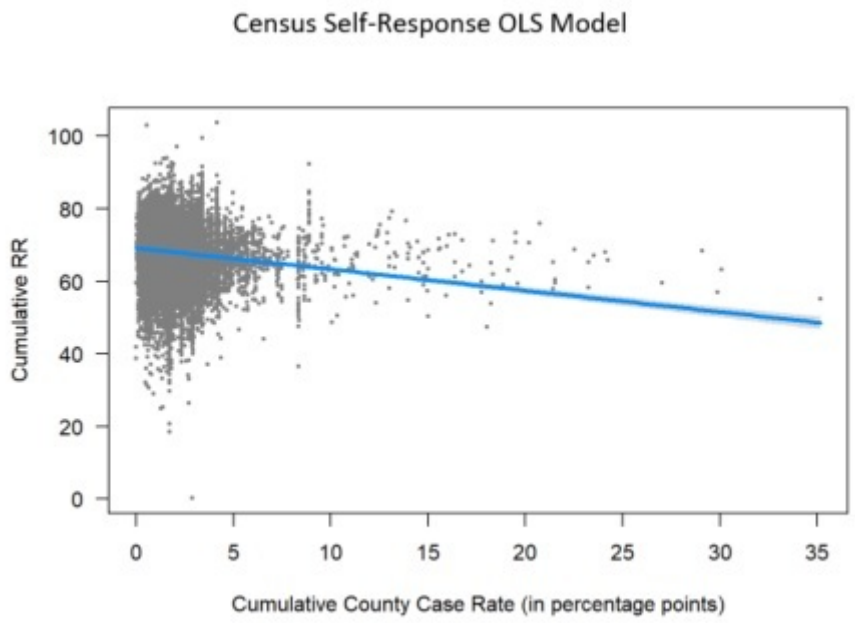

Figure 1: Graph of COVID-19 county-level infection regression coefficient from the Cumulative 2020 Census Self-Response OLS Model

Once we established a significant relationship between infection rates and response, we wanted to understand the importance of COVID-19's predicted effect relative to other factors likely to influence response included in the model. Much of the work on variable "importance" in modeling comes from machine learning classification methods. A general way to gauge variable contribution is to perform machine classification with each variable included and excluded from the model. This process is iterated many times to include many different combinations of variable specifications. With each iteration, some accuracy measure is recorded and attributed to variables included in that iteration. The measures included in models with high measures of accuracy receive high scores for importance and those excluded from models with high accuracy are given poor scores. With enough iterations, it is possible to estimate not only the directionality of each variable's effect, but the relative contribution to predictive accuracy.

We used two methods to measure our linear regression covariate importance. The first was the Filter Variable Importance function from the Caret package in R (Kuhn 2008). Instead of using out-of-sample predictions, this method uses fit measures on the full model data frame. Although it does not have the advantage of testing for overfitting, it does provide a fair comparison of explanatory power for each covariate (Kuhn and Johnson 2013). The method tests many iterations of the model using some or all of the normalized predictors shown in the regression tables. For each iteration, the function records the absolute R^2 value for each variable included. After many iterations, the function records the stacked sum of all the R^2 values to derive the overall score. The final score for each predictor variable is then normalized from 0

Table 3: Contribution of independent variables in the cumulative response rate model.

Variable	Standardized Variable Importance
Median Household Income	0.453982353
% Renter Occupied	0.450266198
% Vacant Units	0.379546222
Cases Rate	0.347104922
% Female House	0.301858715
% Age 18-24	0.293351224
% Black African American in Combo	0.206976666
% Internet First	0.200507293
% White Low Education	0.165951134
% Spanish Speakers	0.126513859
% Age 65 +	0.116642308
State Fixed Effects	0.097394113
% Hispanic Alone	0.096724294
% American Indian/ Alaskin Native in Combo	0.089999554
% Asian in Combo	0.054773335
Density	0.029287131

to 1, so they may be compared. A value of 1 indicates the largest possible contribution, while zero indicates the lowest possible contribution. Table 3 shows the relative contribution of each variable in the cumulative response rate model.

Results show that COVID-19 infection rate was among the most important variables in the model. By this importance score measure, the infection rate had a predictive accuracy similar to that of the percent of female-led households in a tract or the percent of vacant households. Other variables, such as tract percent identifying as Asian alone or in combination with another race, fell low on the relative importance score ranking.

The second method used also comes from machine learning, called the Boruta method for feature selection (Kursa and Rudnicki 2010). Intended to assist with model specification, the method is effective unless the estimation method is computationally intensive (Rudnicki, Wrzesień, and Paja 2015). Unlike the previous method, Boruta creates a threshold for variable inclusion. For example, we know tract percent Asian scored lowest in relative importance but is that too low to include that variable in the model? Boruta, using shuffled data as a baseline, determines if a variable contributes enough predictive power to warrant inclusion. Using many iterations of the data, Boruta shuffles the rows for one variable at a time and sees how predictive the model is versus the version where no data are shuffled. It repeats this technique over many different specifications. The method suggests inclusion for variables that have higher predictive scores than 95% of their shuffled versions. Using the cumulative response rate model, the Boruta method suggests inclusion of each variable in the model, including the state fixed effects (results not shown). Based on results

from these two additional tests, we conclude that our models are reasonably robust in explaining variation in tract response rates and that infection rates played a significant part.

Discussion

In this article, we explore self-response rates to the 2020 Decennial Census and whether an unprecedented environmental event—the COVID-19 virus pandemic—influenced participation in the census. The pandemic upended daily life, with quarantines, business and transportation closures, supply chain disruptions, and the like. In addition, in densely populated cities with high infection rates such as New York City, some residents exited the city seeking areas with lower infection rates. With the first mailing flight arriving at households in mid-March, the Census Bureau was concerned the timing of the pandemic would be particularly disruptive with some residents temporarily displaced. In addition, reporting on the pandemic overtook many media outlets, potentially overshadowing the carefully constructed communication campaign designed to raise awareness, educate the public, and encourage nationwide participation in the census. These events culminated in an unprecedented 2020 Census environment that was no longer "business as usual."

Our analysis is caveated with several limitations. First, because the COVID-19 infection rates are at the county level, we lost some variability in this measure given our unit of analysis was census tract. Since the variance of COVID-19 was decreased due to the aggregation, the variance of the estimate (error term) was, by definition, larger (King, Keohane, and Verba 1994). However, since the model contained over 50,000 observations that spanned over 3,000 counties, the loss in variation on the covariate should have had a minimal effect on the hypothesis test according to the Law of Large Numbers (Finlay and Agresti 1986). Because our unit of analysis was census tract, there was additional between-household variability within tracts not reflected in our models. Also, our measures were cumulative and not a true time series. That means some tracts had a surge in COVID-19 cases during the most crucial portion of the campaign (March), while others experienced it during later "pushes" aimed to count hard-to-survey areas. In addition, community resources and efforts to promote the census varied by state and county, but this exogenous variable is difficult, if not impossible, to operationalize and include in the models (ICF 2012). Finally, with the large number of census tracts under study (over 58,000), the statistical power was large, with most predictor variables being statistically significant. Consequently, we applied several techniques to better understand the explanatory power of each predictor variable.

Our results suggest that even after controlling for variables associated with hard-to-survey populations (e.g., percent female-headed households, percent renter households, percent White persons with less than college education) we found that the higher the county-level rate of COVID-19 infections, the

lower the tract-level self-response rates. We offer several hypotheses to this finding. First, in areas with extremely high infection rates, it is likely the topic dominated media reports, drowning out and/or reducing earned media that might otherwise have helped advertise the 2020 Census. Second, the pandemic caused changes in media consumption, with fewer people performing out-of-home activities such as attending movies or using mass transit, and more consuming in-home media such as Netflix and cable television. In areas with high infection rates, these changes were likely magnified, potentially weakening the impact of a paid advertising campaign designed to increase awareness and ultimately, response. Third, areas with high infection rates undoubtedly experienced high anxiety and uncertainty as a result of health concerns, job losses, school closures, fear of eviction, and other negative physical and mental health outcomes. As a result, the task of completing the census may have simply become a lower priority.

Nancy Bates

U.S. Census Bureau (retired)

batesnancy5@gmail.com

Joseph Zamadics

PSB Insights

jzamadics@psbinsights.com

Submitted: September 04, 2020 EDT, Accepted: January 26, 2021 EDT

REFERENCES

Bunks, Dena, and Jenny Rough. 2020. "List of Coronavirus-Related Restrictions in Every State." AARP. https://www.aarp.org/politics-society/government-elections/info-2020/coronavirus-state-restrictions.html.

COVID-Local. 2020. "A Frontline Guide for Local Decision-Makers." https://www.covidlocal.org/guide/.

Erdman, Chandra, and Nancy Bates. 2017. "The Low Response Score (LRS): A Metric to Locate, Predict, and Manage Hard-to-Survey Populations." *Public Opinion Quarterly* 81 (1): 144–56. https://doi.org/10.1093/poq/nfw040.

Finlay, Barbara, and A. Agresti. 1986. *Statistical Methods for the Social Sciences*. San Francisco, CA: Dellen.

Groves, Robert M., and Mick P. Couper. 1998. *Nonresponse in Household Interview Surveys*. Newark, NJ: John Wiley & Sons, Inc. https://doi.org/10.1002/9781118490082.

ICF. 2012. "2010 Census Evaluation of National Partnership Research Report." 2010 Census Planning Memorandum Series, No. 196. https://www2.census.gov/programs-surveys/decennial/2010/program-management/5-review/cpex/2010-memo-196.pdf?#.

Johnson, Timothy P., Young IkB Cho, Richard T. Campbell, and Allyson L. Holbrook. 2006. "Using Community-Level Correlates to Evaluate Nonresponse Effects in a Telephone Survey." *Public Opinion Quarterly* 70 (5): 704–19. https://doi.org/10.1093/poq/nfl032.

King, Gary, Robert O. Keohane, and Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton, NJ: Princeton University Press. https://doi.org/10.1515/9781400821211.

Knittel, Christopher R., and Bora Ozaltun. 2020. "What Does and Does Not Correlate with COVID-19 Death Rates." *CEEPR Working Paper* 2020–009 (June). https://doi.org/10.3386/w27391.

Kuhn, Max. 2008. "Building Predictive Models in R Using the Caret Package." *Journal of Statistical Software* 28 (5): 1–26. https://doi.org/10.18637/jss.v028.i05.

Kuhn, Max, and Kjell Johnson. 2013. "Measuring Predictor Importance." In *Applied Predictive Modeling*, 463–85. New York: Springer. https://doi.org/10.1007/978-1-4614-6849-3_18.

Kursa, Miron B., and Witold R. Rudnicki. 2010. "Feature Selection with the Boruta Package." *Journal of Statistical Software* 36 (1): 1–13.

Lee, Jasmine C., Sarah Mervosh, Yuriria Avila, Barbara Harvey, and Alex Leeds Matthews. 2020. "See How All 50 States Are Reopening (and Closing Again)." *New York Times*, 2020. https://www.nytimes.com/interactive/2020/us/states-reopen-map-coronavirus.html.

Liu, Tao, Wenjia Liang, Haojie Zhong, Jianfeng He, Zihui Chen, Guanhao He, Tie Song, et al. 2020. "Risk Factors Associated with COVID-19 Infection: A Retrospective Cohort Study Based on Contacts Tracing." *Emerging Microbes & Infections* 9 (1): 1546–53. https://doi.org/10.1080/22221751.2020.1787799.

Moore, Dasia, and Kay Lazar. 2020. "Experts Urge Rollback of Reopening as COVID-19 Cases Rise in Massachusetts: A Northeastern University Epidemiologist Is Advocating for a Return to Phase 2 of the State's Plan." *Boston Globe*, August 3, 2020. https://www.bostonglobe.com/2020/08/03/nation/amid-rise-covid-19-cases-experts-eye-roll-back-reopening-mass/.

National Public Radio. 2020. "What's Behind States' Differing Approaches To Reopen Economies?" *Morning Edition*. https://www.npr.org/2020/05/04/849927408/whats-behind-states-differing-approaches-to-reopen-economies.

Orgera, Kendal, Daniel McDermott, and Matthew Rae. 2020. "Urban and Rural Differences in Coronavirus Pandemic Preparedness." https://www.kff.org/coronavirus-covid-19/issue-brief/urban-and-rural-differences-in-coronavirus-pandemic-preparedness/.

Rudnicki, Witold R., Mariusz Wrzesień, and Wiesław Paja. 2015. "All Relevant Feature Selection Methods and Applications." In *Feature Selection for Data and Pattern Recognition*, edited by U. Stańczyk and L. Jain, 11–28. Studies in Computational Intelligence 584. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-662-45620-0_2.

University of Minnesota. 2020. "COVID-19 Hospitalizations Analysis Shows Disparities across Racial and Ethnic Groups." https://www.sciencedaily.com/releases/2020/08/200817150443.htm.

U.S. Census Bureau. 2020. "Statement from U.S. Census Bureau Director Steven Dillingham: Delivering a Complete and Accurate 2020 Census Count." Press Release, AUGUST 03, 2020, RELEASE NUMBER CB20-RTQ.23. https://www.census.gov/newsroom/press-releases/2020/delivering-complete-accurate-count.html.

U.S. Chamber of Commerce. 2020. "State-by-State Business Reopening Guidance." https://www.uschamber.com/article/state-by-state-business-reopening-guidance.

Weise, Elizabeth. 2020. "Wearing a Mask Doesn't Just Protect Others from COVID, It Protects You from Infection, Perhaps Serious Illness, Too." *USA TODAY*, 2020. https://www.msn.com/en-us/news/us/wearing-a-mask-doesn-t-just-protect-others-from-covid-it-protects-you-from-infection-perhaps-serious-illness-too/ar-BB16LciZ.

Appendix

List of Variables Included in the Model

DEPENDENT VARIABLE

Cumu. RR: Cumulative response rate. Percent of eligible households in a tract that had self-responded to the census as of July 25, 2020.

INDEPENDENT VARIABLES

Cases-rate: The tract's county cumulative positive tests (as of July 25, 2020) divided by the 2017 reported county population (Source: CDC).

% Hispanic alone: The PDB reported percentage of persons in a tract reporting to be Hispanic alone.

% Black African American in combo: The PDB reported percentage of persons in a tract reporting to be Black/African American alone or in combination with another race.

% Asian in combo: The PDB reported percentage of persons in a tract reporting to be Asian alone or in combination with another race.

% American Indian or Alaska Native in combo: The PDB reported percentage of persons in a tract reporting to be American Indian/Alaska Native alone or in combination with another race.

% White low education: The PDB reported percentage of persons in a tract reporting to be White alone or in combination with another race and no college education.

% Spanish speakers: The PDB percent of households in a tract reporting to speak Spanish, with no one aged 14 or over speaking English very well.

% Vacant units: The PDB percent of households in a tract reported as vacant.

% Renter occupied households: The PDB percent of occupied households in a tract not owner occupied (including those rented for cash and those occupied without rent payment).
Median HH income: The PDB Census tract median household annual income.

% Age 65+: The PDB percent of persons in a tract reporting as aged 65+.

% Age 18–24: The PDB percent of persons in a tract reporting as aged 18–24.

% Female House: The PDB percent of households in a tract reporting as female headed, no spouse present.

Density: ([The PDB 2017 population in a tract /100] / the PDB tract land area in square miles).

Internet Choice: If tract has distinction as Internet Choice or Internet First (Internet choice set as reference segment). Source: Census Bureau.

States: Fixed effects (factor variable) for each state including DC (Alabama set as reference state).