Tolliver, Kevin, Jason Fields, Renee Stepler, and Abby Williams. 2023. "Case Prioritization in the SIPP: A Five-Year Review." *Survey Practice* 16 (1). https://doi.org/10.29115/SP-2023-0010.

ARTICLES

Case Prioritization in the SIPP: A Five-Year Review

Kevin Tolliver, Jason Fields, Renee Stepler, Abby Williams

Keywords: case prioritization, adaptive design, responsive design, field management, r-indicators, nonresponse bias

https://doi.org/10.29115/SP-2023-0010

Survey Practice

Vol. 16, Issue 1, 2023

Case prioritization is a concerted effort to achieve high data quality during data collection by reallocating scarce resources to the sample cases that need them the most. Our research, conducted as part of the Survey of Income and Program Participation (SIPP), examines the effect of case prioritization over the changing landscape of this survey. Our findings suggest that since instituting case prioritization, field interviewers exert more effort on higher priority cases, and continual centralized monitoring and intervention have led to improvements in data quality measures, in particular R-indicator and coefficient of variation (CV) of response propensities.

Introduction

The tools, procedures, and designs for modifying strategies during data collection have proven to be valuable for the Survey of Income and Program Participation (SIPP) despite continual decline in overall response rates. The SIPP is a complex survey whose potential value becomes more evident with the programmatic and policy questions raised during challenging times. While the survey environment for longitudinal household panel surveys has always been challenging, addressing increasing nonresponse, focusing on data quality measures, and being sensitive to the respondent and interviewer burden are even more critical when the collection is dramatically interrupted.

The SIPP is a longitudinal government survey that collects information on a variety of topics, which allows for the study of interaction between tax, transfer, and other government and private polices, with multiple social and demographic characteristics. The SIPP is a computer-assisted person interview (CAPI) survey collected primarily in person. Respondents in the first wave remain in the sample for the next three data collection years regardless of whether they move to another address. The first wave nonrespondents are dropped from the sample but are replaced with a new sample, so the total sample size remains at approximately 53,000 households. Further information of the survey, its sample design, content, and uses are available on http://www.census.gov/sipp.

Despite challenges of uncertainty of funding for SIPP in 2018, a 35-day federal furlough during the peak hiring and training for SIPP in 2019, followed by the incredible impacts of the coronavirus pandemic in 2020 and 2021, case prioritization efforts have allowed the SIPP program to adapt to unexpected conditions and to make well-informed decisions throughout data collection aimed at attaining respondents that are representative of the population. This research examines the effect of case prioritization on representation during the last five years.

Background

Case prioritization refers to targeting a subset of cases with pre-identified data collection features that are different from the typical features applied to the nontargeted population. SIPP targets cases that belong to subgroups with historically low response rates to help produce a representative respondent population. Interviewers are instructed to give targeted cases "first attention" and to "make at least one attempt per week." Though the instructions are intentionally left vague, so that interviewers have a little flexibility regards how they alter their contact strategy, it typically means more contact attempts and more time spent on these cases, which reduces contact attempts and attention on other remaining cases. Increasing response in under-represented populations while decreasing response in over-represented populations reduces the variance in response rates of differing populations.

Case prioritization is a common way of allocating scarce resources to prioritized cases (Schouten, Peytchev, and Wagner 2017). Prior studies have shown that case prioritization has had varying effects on data quality. The following surveys applied case prioritization differently to attend to their goals leading to a variety of outcomes. The National Survey of Family Growth (NSFG), a CAPI in-person survey, conducted multiple prioritization experiments with differing goals where targeted cases were instructed to receive more calls. Their prioritization rarely improved survey outcomes, such as response rates (Tourangeau et al. 2017). Statistics Canada capped the number of calls on nontargeted cases with the goal of increasing representativeness in their surveys. They found that while their prioritization did not lead to much increase in representativeness, it did lead to a reduction in interviewing hours (Tourangeau et al. 2017). The case prioritization did not affect the response rate which meant a decrease in interviewing hours per response as well.

The Community Advantage Panel Survey (CAPS), a CAPI in-person survey, offered larger interviewer incentives for completing low response propensity cases with the goal of increasing overall response. This reduced variation in the response propensities but did not reduce nonresponse bias as intended (Peytchev et al. 2010). The High School Longitudinal Study, a multi-mode survey with web and phone modes as the initial mode, targeted the lowest propensity quartile with an increase in in-person visits, which lowered the average relative bias by 0.4% (Schouten, Peytchev, and Wagner 2017). The National Survey of College Graduates (NSCG), a web and phone-mode government survey, found evidence that case prioritization in the form of a system of continuous monitoring and intervening protocols, can improve R-indicators (Coffey, Reist, and Miller 2020).

Past research on the SIPP found that prioritizing cases led to smaller attrition bias in unweighted key estimates and helped increase retention of movers through a pilot study first conducted in 2016 followed by a larger interviewer-level experiment in 2017 (Tolliver et al. 2019).

The effect of case prioritization in SIPP is reliant on interviewer compliance, intervention effectiveness, and proper targeting. It is documented that interviewers do not always comply with prioritization instructions (Nagle and Walejko 2018; Walejko and Miller 2015; Walejko and Wagner 2015). With full interviewer compliance, the effect still relies on the effectiveness of the intervention. We assume that the interviewer altering their contact strategy leads to higher response propensities. If, however, the altered contact strategy has no impact on responding, then our prioritization has no effect. Lastly, if we have interviewer compliance and an effective intervention but incorrectly target cases, then we may introduce bias instead of reducing bias.

Methods

The primary prioritization goal in SIPP is to obtain respondents representative of the United States population. Through continuous monitoring and intervention, SIPP instructs interviewers to reallocate their efforts to cases that may have the highest impact on data quality. Every case on the active workload has a priority status of high, medium, or low. Using a combination of known and unknown demographic and geographic predictors (see Appendix A2), SIPP identifies which cases are most at risk of being under-represented and designates them as high priority. Cases at risk of being overrepresented and cases that have been worked appropriately with low chances of responding are designated as low priority. All other cases are designated as medium priority.

Interviewer Compliance

Interviewers are instructed to give high priority cases first attention each day they work. They are also instructed to work their medium priority cases as they normally would and to not make any attempts on low priority cases until they have made sufficient effort on their high and medium priority SIPP cases.

Typically, at the start of data collection, about 20% of the workload is designated high priority and 80% of the workload is designated medium priority. Through the 2021 data collection, there has never been any low priority cases at the start of data collection.

The priority status for a case may change as often as every two weeks, though the priority status of a case usually does not change more than once throughout the entire data collection. SIPP continually monitors paradata, interviewing costs, and the distribution of respondents when considering priority changes. For simplicity, we categorize these prioritizations as early-stage (s), mid-stage (m), and late-stage (l) prioritizations.

The interviewers are required to communicate with the case management systems regularly to see the most up-to-date priority. When a case becomes high priority, interviewers are instructed to make a contact attempt within one week. Late in data collection, some low priority cases are stopped to ensure interviewers focus on their higher priority cases. Once a case is stopped, no additional attempts are allowed. Cases that are stopped are assumed to be eligible nonrespondents.

We assess if interviewers are complying with instructions by observing the percent of cases that had attempts within the first seven days of the initial prioritization and the differences in weekly attempts per case for high, medium, and low priority cases.

Intervention Effectiveness

Representativeness-indicators "R-indicators" measure representation by observing one minus the variance in the response propensities given demographic and geographic characteristics (Schouten, Cobben, and Bethlehem 2009). Smaller variances are indicative that the demographic and geographic characteristics contribute equally to nonresponse, and thus, we assume larger R-indicators are likely to have smaller nonresponse biases.

A secondary goal of case prioritization in the SIPP is not to sacrifice overall response while in pursuit of the most representative population possible. The coefficient of variation (CV; De Heij, Schouten, and Shlomo 2015) in the response propensities measures representation in relation to response. In contrast to R-indicators, smaller CVs are likely to have smaller nonresponse biases. Because this metric has the mean response propensity as a divisor, it is assumed that smaller CVs are more likely to have smaller mean squared errors.

We assess if the intervention is effective by comparing R-indicators and CVs of the treated sample, where there was a mix of high, medium, and low priority cases, to the nontreated sample where all cases were displayed as medium priority. The R-indicator and CV are computed separately for the Wave 1 and Wave 2+ samples since the Wave 2+ sample has the benefit of having prior wave information as predictors of the R-indicator and CV.

While both the 2017 and 2018 data collections were experimental years,[1] where we can simply compare the treatment to the control, the 2019-2021 data collections are thought as nonexperimental years because there has been no formal control group. This means our comparisons of the treated sample to the non-treated sample are observational. This paper estimates what the R-indicator and CV might have been without prioritization by matching high

[1] The 2017 calendar year only consisted of Wave 4 data and the 2018 calendar year only consisted of Wave 1 data.

and low priority cases to similar medium priority cases (not made high or low for evaluation purposes), then adjusting the propensities by an estimated effect. The steps for the analyses for 2019–2021:

1. Compute observed R-indicator, mean response propensity, and CV as

$$\widehat{R} = 1 - 2s\left(\widehat{\rho}\right)$$

$$\bar{\widehat{\rho}} = n^{-1}\sum_{i=1}^{n}\widehat{p}_i$$

$$\widehat{CV} = s(\widehat{\rho})/\bar{\widehat{\rho}}$$

   where $s(\widehat{\rho})$ is the standard deviation of response propensities.

2. Match any case that was high priority, low priority, or stopped using greedy nearest-neighbor propensity matching using demographic, geographic, calendar year, stage, and interviewer caseload information. (SAS Documentation; Stuart 2010) "Similar" in the context of this paper refers to medium priority cases that were matched to the high or low priority cases.

3. Estimate the response propensity effect $\widehat{\beta_{p,t}}(\rho)$ by

$$\widehat{\beta_{p,t}}(\rho) = \left(\frac{I}{I+NI}\right)_{Treated\ data} - \left(\frac{I}{I+NI}\right)_{Matched\ data}$$

   where I is number of interviews, NI is the number of eligible noninterviews, p is priority (H,M,L), t is the time period when the case was made priority p (e= early-stage [first 4 weeks], m=mid-stage [5th week to 7th to last week], l=late stage [last 6 weeks]).

4. Re-estimate propensity scores as

$$\widetilde{\rho}_i = \widehat{\rho}_i - \widehat{\beta_{p,t}}(\rho)$$

5. Compute the expected R-indicator, mean response propensity, and CV if no prioritization was used as

$$\widetilde{R} = 1 - 2s\left(\widetilde{\rho}\right)$$

$$\bar{\widetilde{p}} = n^{-1}\sum_{i=1}^{n}\widetilde{p}_i$$

$$\widetilde{CV} = s(\widetilde{\rho})/\bar{\widetilde{\rho}}$$

6. Estimate the R-indicator, CV, and response rate effect, ($\widehat{\beta}(R)$, $\widehat{\beta}(CV)$, $\widehat{\beta}(\rho)$) respectively, by

Table 1. Estimated Case Prioritization Effect on Contact Attempts During Different Periods in Data Collection with Standard Errors.

Data Collection	High	Medium similar H	Low	Medium Similar L	Medium Other
Early: First 4 weeks	0.428 (0.0049)	0.276 (0.0041)	-	-	0.231 (0.0010)
Mid: Middle 7 to 11 weeks[a]	0.305 (0.0038)	0.266 (0.0030)	0.316 (0.0204)	0.279 (0.0021)	0.237 (0.0010)
Late: Last 6 weeks	0.249 (0.0024)	0.236 (0.0021)	0.199 (0.046)	0.270 (0.0028)	0.178 (0.0007)

[a]The number of middle weeks is the difference of the total length of data collection in weeks minus 10.

$$\widehat{\beta}(R) = 100 \times (\widehat{R} - \widetilde{R}) \div \widetilde{R}$$
$$\widehat{\beta}(CV) = 100 \times (\widehat{CV} - \widetilde{CV})$$
$$\widehat{\beta}(\rho) = 100 \times (\overline{\widehat{\rho}} - \overline{\widetilde{\rho}}) \div \overline{\widetilde{\rho}}$$

Results

The prioritization led to increased effort on higher priority cases and decreased effort on lower priority cases, though neither as much as intended. Overall, there were more attempts on high priority cases. The difference in attempts among high priority and all other cases was larger at the start of data collection but became smaller throughout data collection. Even though the instructions state "make at least one attempt every week," fewer than half of the high priority cases had one attempt within the first seven days of becoming high priority. Overall, there were not fewer attempts on low priority cases, but there were fewer attempts on low priority cases compared to similarly matched medium priority cases. Low priority cases that were stopped ultimately led to more attempts on other remaining cases. For every 1,000 cases stopped, the remaining workload receives about +0.05 more weekly attempts. In addition, fewer than half of the initial high priority cases had attempts within the first seven days of data collection, but many more were started the first week compared to their medium priority counterparts (H = 37% vs. M = 27%). The median first attempt for cases prioritized at the start of data collection is day 22 compared with day 30 for all other cases.

Table 1 illustrates how the priority status affected the number of weekly contact attempts per case during different periods in data collection. This is calculated as the total attempts divided by the total number of longitudinal case observations[2] by priority category during that period.

[2] The paradata datasets are longitudinal observations, with an observation for each week; however, the case still requires more work for data collection to be complete. This means each week, there are fewer total observations.

Table 2. Estimated Case Prioritization Effect on Response Propensities with Standard Errors.

Data Collection	Wave 1 High	Wave 2+ High	Wave 1 Low	Wave 2+ Low	Wave 1 Low and Stopped	Wave 2+ Low and Stopped
Early: First 4 weeks	+0.10 (0.007)	+0.20 (0.009)				
Mid: Middle 7 to 11 weeks	+0.02 (0.011)	+0.02 (0.015)				
Late: Last 6 weeks	+0.05 (0.011)	+0.00 (0.029)	-0.14 (0.009)	-0.09 (0.010)	-0.21 (0.006)	-0.11 (0.001)

Source: U.S. Census Bureau, Survey of Income and Program Participation 2017-2021, Project No. P-7529920, Approval CBDRB-FY21-POP001-0128

Table 3. Estimated Case Prioritization Effect on the Data Quality Indicators: R-indicator (R), Coefficient of Variation (CV), and Mean Overall Response Propensity (ρ).

	Wave 1			Wave 2+		
Year	$\widehat{\beta}(\mathbf{R})$	$\widehat{\beta}(\mathbf{CV})$	$\widehat{\beta}(\rho)$	$\widehat{\beta}(\mathbf{R})$	$\widehat{\beta}(\mathbf{CV})$	$\widehat{\beta}(\rho)$
2017	NA	NA	NA	+3.0	-1.1	-1.2
2018	+3.6	-1.1	-2.4	NA	NA	NA
2019	+3.1	+17.2	-14.0	+6.6	-4.8	+0.8
2020	+1.3	-7.6	+4.1	+8.3	-6.1	+0.2
2021	+1.4	-4.2	+3.2	+5.3	-5.3	+2.2

$Where\ \widehat{\beta}(R) = 100 \times (\widehat{R} - \widetilde{R}) \div \widetilde{R}, \widehat{\beta}(CV) = 100 \times (\widehat{CV} - \widetilde{CV}), and\ \widehat{\beta}(\rho) = 100 \times (\widehat{\overline{\rho}} - \widetilde{\overline{\rho}}) \div \widetilde{\overline{\rho}}$

Source: U.S. Census Bureau, Survey of Income and Program Participation 2017-2021, Project No. P-7529920, Approval CBDRB-FY21-POP001-0128

Interviewers with prior SIPP experience and/or with larger SIPP workloads were 10% to 20% more likely to work their high priority cases within the first seven days of becoming high priority. Table 2 summarizes how the differing effort affected the response propensities.

The observed Wave 1 R-indicator was on average 2.1% larger than the expected Wave 1 R-indicator if no prioritization was used, while having a CV that was on average 4.3 percentage points smaller than the expected Wave 1 CV. The observed Wave 2+ R-indicator was on average 5.8% larger than the expected Wave 2+ R-indicator, while having a CV that was on average 4.3 percentage points smaller than the expected Wave 2+ CV. The observed mean response propensity and estimated mean response propensity were nearly the same. Though these results may not be statistically significant, they are generally favorable. Table 3 gives a high-level summary of case prioritization spanning multiple years.

Discussion

As noted in the introduction, each year the survey has faced a different set of challenges making the estimation step in propensity matching imperfect. Funding uncertainty, government shutdowns, and a global pandemic have made it so that no two years are exactly alike. See Appendix A1 for more details. This means across years, every case can be matched, but there is variability from one year to the next, and while we believe that data has benefited from case prioritization, our method for evaluating is a limitation. All data are subject to errors arising from a variety of sources.

Despite this limitation, the belief is that the case prioritization benefits overall representation. Furthermore, by design, our case prioritization method has helped the survey's nonresponse adjustments because we use known covariates that are closely associated with the survey's nonresponse adjustments to target cases. Perhaps another method of evaluating the prioritization's effectiveness is estimating the prioritization's effect on the nonresponse adjustments.

Summary

Though response rates have continued to suffer in SIPP in the wake of numerous challenges, we believe that prioritization has helped mitigate some issues in data quality. The prioritization effects are not large but consistently positive. There has been a consistent increase in R-indicator, and outside of the 2019 Wave 1 sample where it was decided to stop most cases to focus on the 2018 Wave 2 sample, there has been a consistent decrease in CV. The prioritization improved data quality indicators (R = +2.1%, CV = -4.3 percentage points) for Wave 1 data and (R = +5.8%, CV = -4.3 percentage points) for Wave 2+ data. In the most recent years, where we have leveraged information from prior years, we believe that the prioritization helped the overall response rate (+3.6 for Wave 1, +1.2 for Wave2+).

This paper provides evidence that dynamic case prioritizations can be a tool to improve data quality. While there is evidence of positive effects, the results leave room for improvement. As SIPP plans to continue use of case prioritization, the survey program will test new strategies to increase interviewer compliance, explore administrative data to improve estimation of response propensities, and consider the tradeoffs between the increase in representation and costs.

Disclaimer

Any opinions and conclusions expressed herein are those of the author(s) and do not reflect the views of the U.S. Census Bureau.

Acknowledgements

We extend a thank you to Bennett Adelman, Carolyn Pickering, and Rachel Horwitz for the review of this paper.

Submitted: February 09, 2022 EST, Accepted: July 07, 2023 EST

REFERENCES

Coffey, Stephanie, Benjamin Reist, and Peter V. Miller. 2020. "Interventions On-Call: Dynamic Adaptive Design in the 2015 National Survey of College Graduates." *Journal of Survey Statistics and Methodology* 8 (4): 726–47. https://doi.org/10.1093/jssam/smz026.

De Heij, V., B. Schouten, and N. Shlomo. 2015. *RISQ 2.1 Manual. Tools in SAS and R for Computation of R-Indicators and Partial R-Indicators*. https://hummedia.manchester.ac.uk/institutes/cmist/risq/RISQ-Deliverable-12-1.pdf.

Nagle, Amanda, and Gina Walejko. 2018. "Exploring Reminder Calls Intended to Increase Interviewer Compliance with Data Collection Protocols." *Survey Practice* 11 (2): 1–17. https://doi.org/10.29115/sp-2018-0022.

Peytchev, A., S. Riley, J. Rosen, J. Murphy, and M. Lindblad. 2010. "Reduction of Nonresponse Bias in Surveys through Case Prioritization." *Survey Research Methods* 4: 21–29.

Schouten, Barry, F. Cobben, and J. Bethlehem. 2009. "Indicators for the Representativeness of Survey Response." *Survey Methodology* 35: 101–13.

Schouten, Barry, Andy Peytchev, and James Wagner. 2017. *Adaptive Survey Design*. Boca Raton, Florida: CRC Press. https://doi.org/10.1201/9781315153964.

Stuart, E. 2010. "Matching Methods for Causal Inference: A Review and a Look Forward." *Stat Sci* 25 (1): 1–21.

Tolliver, K., J. Fields, S. Coffey, and A. Nagle. 2019. "Combatting Attrition Bias in the Survey of Income and Program Participation." In *Proceedings of the Joint Statistical Meetings 2019, Denver, Colorado. July 27-August 1, 2019.*

Tourangeau, Roger, J. Michael Brick, Sharon Lohr, and Jane Li. 2017. "Adaptive and Responsive Survey Designs: A Review and Assessment." *Journal of the Royal Statistical Society Series A: Statistics in Society* 180 (1): 203–23. https://doi.org/10.1111/rssa.12186.

Walejko, Gina, and Peter Miller. 2015. "The 2013 Census Test: Piloting Methods to Reduce 2020 Census Costs." *Survey Practice* 8 (6): 1–8. https://doi.org/10.29115/sp-2015-0030.

Walejko, Gina, and J. Wagner. 2015. "Challenges to Innovation in Face-to-Face Surveys Posed by Interviewer Noncompliance." In *The 70th Annual Conference of the American Association of Public Opinion Research, Hollywood, FL, May 14–17, 2015.*

Appendices

Appendix A1

This section describes how the challenges faced by SIPP affected the case prioritization strategy.

The 2018 data collection was the first Wave 1 sample of SIPP that used case prioritization. There was uncertainty that the survey would receive full funding to field the entire sample. This led to changing the survey design to monthly panels instead of one panel. This affected the case prioritization because there was at most one month to be worked. While on average it takes about two weeks to fully resolve a case, many of the high priority cases require longer than a month.

The 2019 data collection was the first sample that had the Wave 1 sample and Wave 2+ sample in the field concurrently. The 35-day government shutdown occurred during the peak time for onboarding interviewers, forcing hiring numbers to be extremely low. The shutdown impacted progress for both the Wave 1 and Wave 2 sample, but disproportionately impacted Wave 1. SIPP decision makers elected to best preserve longitudinal panel data and sacrifice the Wave 1 sample.

The 2020 data collection was heavily impacted by the coronavirus pandemic, which forced a shift from in-person interviewing to 100% telephone interviewing midway through data collection. This was particularly challenging for Wave 1 households that did not have any telephone information provided. While Wave 2+ households had telephone numbers from the prior wave, Wave 1 households not contacted before March 19th relied on three best possible telephone numbers provided by the administrative records team.

The 2021 data collection had the in-person attempt restrictions lifted for much of the survey, but there was still reluctance by interviewers and interviewees to conduct in-person interviews.

Appendix A2

The variables used to calculate propensity scores were a combination of prior geographic data and demographic data coming from the planning database, prior wave interviews, and post-data collection edits.

Table A1. Variables used for Propensity Scores.

New Wave 1 Sample	Returning Wave 2+ Sample
Census Region	Census Region
Metropolitan Statistical Area (MSA)/Not MSA	Metropolitan Statistical Area (MSA)/Not MSA
Poverty Stratum	Poverty Stratum
Number of people in household	Number of people in household
Race	Race
Tenure	Tenure
Percent of Female No Husband in Block	Female Householder
Urban/Rural	Urban/Rural
Percent of Non-English speakers in Block	Percent of Non-English speakers in Block
Percent of No High School Diploma in Block	Percent of No High School Diploma in Block
Percent of Block Receiving Public Assistance	Percent of Block Receiving Public Assistance
Percent of Block Aged 65+	Age of Eldest Person in household
	Age of Youngest Person in household