HRS | HEALTH AND RETIREMENT STUDY
A Longitudinal Study of Health, Retirement, and Aging
Sponsored by the National Institute on Aging

SRC Data Quality Profile

Documentation Report
HRS 2022 Panel

Heather Schroeder, HRS Statistician Staff Specialist, University of Michigan
Chendi Zhao, Doctoral Candidate (Survey and Data Science), University of Michigan
Sergio Martinez, Doctoral Candidate (Survey and Data Science), University of Michigan
Brady T. West, HRS Associate Director, University of Michigan

Survey Research Center
Institute for Social Research
University of Michigan
Ann Arbor, Michigan

April 2026

Funding
The Health and Retirement Study is funded by a grant from the National Institute on Aging (U01 AG009740) with supplemental support from the Social Security Administration. HRS is conducted by the University of Michigan.

Suggested Citation
Schroeder, H., Zhao, Z., Martinez, S. & West, B. (2026). SRC Data Quality Profile. University of Michigan. https://hrs.isr.umich.edu/publications/biblio/15957

SRC Data Quality Profile: HRS 2022 Panel Introduction

The purpose of this Data Quality Profile (DQP) is to objectively describe the quality of the HRS 2022 panel survey from the Total Survey Error (TSE) perspective. This profile consists of 23 criteria that combine quantifiable indicators of data quality and brief text descriptions that provide an overview of the data collection process. These indicators are grouped into three broad domains: the nonresponse error / selection bias domain, the measurement error domain, and the sampling error domain.

The HRS 2022 Panel DQP includes HRS panelists who were enrolled in the study prior to 2022. During the 2022 HRS data collection, a new cohort (Early Gen X, those born between 1966 and 1971 and their younger spouses) and a Minority Older Cohort were simultaneously being recruited to join the HRS. Newly recruited persons empaneled to the Early Gen X and Minority Older Cohort are not included in this report. In some cases, as HRS panelists age, a proxy is utilized to help panelists participate in the survey. Proxy responses are included throughout the DQP unless noted otherwise. When an HRS panelist dies, an “Exit” or “Post-Exit” interview is conducted with a proxy to obtain a final set of information. “Exit” and “Post-Exit” proxy interviews are not included in the DQP.

Creating the first HRS DQP required a relatively large setup cost. For the initial HRS DQP for the 2022 panel survey, three people spent roughly 40 hours per month for about two and a half months to complete all 23 items. A total of 10 items were created or written before the development of the first DQP, and therefore only required minimal effort to incorporate into the report. The other 13 items were developed from scratch and therefore took more time. Subsequent DQPs created for other waves of HRS data collection can take advantage of the code that was created for this initial profile and therefore should take significantly less time to create. Other efforts to create this type of DQP will also require the prior identification of a handful of “key variables” that are, in the eyes of a given study, of significant value to the study’s stakeholders and researchers. This process for the HRS 2022 panel was completed before this report was finalized.

At the time of this writing, calibration adjustments are still being finalized for the 2022 HRS to incorporate individuals from the EGENX and MOC recruitment. Therefore, all HRS estimates in this report are subject to change upon our receipt of the final weights. The HRS weights for panel respondents used throughout the DQP are considered “interim” or “early-release”.

Nonresponse Error / Selection Bias

1. AAPOR (RR6) Response Rate

The AAPOR RR6 for the HRS 2022 panel was 67.9%. (Reference)

2. Overall / Partial R-Indicators

- **Household level response propensity model**

The R-indicator (Schouten et al., 2009) provides a measure of the representativeness of a realized sample based on the variation in predicted response propensities across different subgroups. To obtain an R-indicator for the 2022 wave of the Health and Retirement Study, we modeled household-level response propensities using a sample restricted to panel members who responded in 2020. A household was classified as responding in 2022 if at least one eligible household member completed an interview according to the 2022 tracker file.

Household-level predictors were taken from the 2020 HRS data and include cohort membership, Baby Boom versus pre–Baby Boom status, household structure, race and ethnicity of the household reference person, educational attainment, partnership and marital status, presence of functional limitations, and household size. The total number of telephone contact attempts during the 2020 field period was also included as a predictor.

We estimated several alternative response propensity models, including standard logistic regression, elastic net penalized logistic regression, random forests implemented via ranger, neural networks, gradient boosted trees using xgboost, and Bayesian additive regression trees (BART). Model comparison was based on cross-validated area under the ROC curve, using a training-test split. For models requiring tuning, hyperparameter grids were specified and the configuration that maximized the cross-validated AUC was selected.

The model with the highest AUC was selected as the final response propensity model and used to generate predicted household response probabilities for computing the R-indicators.

Model	AUC		Model	AUC
Logistic	0.573		Neural Network	0.775
Elastic Net	0.769		XGBoost	0.788
Random Forest	0.787		**BART**	**0.793**

As shown in the table above, BART provided the strongest predictive performance, with an AUC of 0.793, and was therefore selected as the final response propensity modeling approach.

- **Overall R-indicator**

Based on the resulting household-level predicted response probabilities, the R-indicator for the 2022 wave was 0.682. Given that an R-indicator of 1 is “perfect” and indicates no variability in response propensities for the realized sample, this result indicates moderate variability in response propensities across the sample and suggests that standard nonresponse adjustment procedures applied by HRS remain relevant for repairing potential nonresponse bias in the estimates. We note that there is no standard cut-off for the R-indicator that indicates whether standard nonresponse adjustments would be effective, but that as a measure of representativeness, is does assume that nonresponse is occurring at random conditional on the covariates used to compute the metric.

- **Variable-level (Cohort) unconditional partial R-indicator**

Variable	Partial R-Indicator
Cohort	0.0258

In the small table above, the variable-level unconditional partial R-indicator for cohort is presented. Since unconditional partial R-indicators are bounded above by 0.5, the reported value of 0.0258 indicates that differences in response propensities across cohorts explain only a limited share of the total variability in response propensities. In absolute terms, cohort plays a relatively minor role in driving nonrepresentative response, suggesting balanced response across the cohorts.

- **Category-level unconditional partial R-indicator**

Category	Partial R-Indicator
AHEAD+CODA+HRS	-0.0018
War Babies (WB)	0.0053
Early Baby Boomers (EBB)	0.0129
Mid Baby Boomers (MBB)	0.0078
Late Baby Boomers (LBB)	-0.0202

In the table above, the category-level unconditional partial R-indicators for cohort are presented. These values can range between −0.5 and 0.5. All category-level indicators are close to zero, indicating only mild deviations from representative response across the cohorts. Positive values for the WB, EBB, and MBB cohorts indicate slight overrepresentation, while the negative values for the AHEAD+CODA+HRS and LBB cohorts indicate slight underrepresentation. The LBB cohort had the lowest observed response rate in the 2022 wave (59.7%).

3. Subgroup Variation in Response Propensity

The overall response rate (AAPOR RR6) for the 2022 HRS panel was 67.9%. The table below shows some variability in response propensity by race and study cohort. Non-Hispanic White/Others and the most recent study cohorts had the lowest response propensity. To the extent that these subgroups vary in terms of key HRS measures, slight nonresponse bias might be introduced in estimates based on those measures. Nonresponse adjustments based on these observable subgroups would help to repair this bias.

Response Rates: 2022 HRS Panel	
Overall	**67.9%**
Race/Ethnicity	
Hispanic	67.5%
Non-Hispanic Black	71.1%
Non-Hispanic White/Other	66.8%
Study Cohort	
HRS	71.7%
AHEAD	68.6%
Children of the Depression Era	64.0%
War Babies	74.0%
Early Baby Boomers	72.1%
Mid Baby Boomers	67.2%
Late Baby Boomers	59.7%

Reference

4. References to Benchmarks

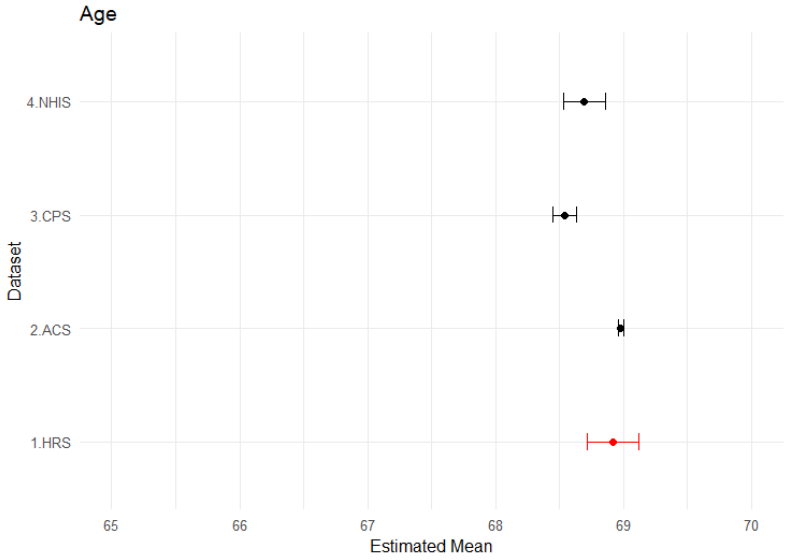
We identified common variables across the HRS, the American Community Survey (ACS), the Current Population Survey (CPS), and the National Health Interview Survey (NHIS), and compared national estimates based on the HRS to estimates based on these other three benchmark surveys. Our analysis used data from all sources in 2022 and was restricted to non-institutionalized individuals aged 57 and older, aligning with the youngest age cohort in HRS in the 2022 panel.

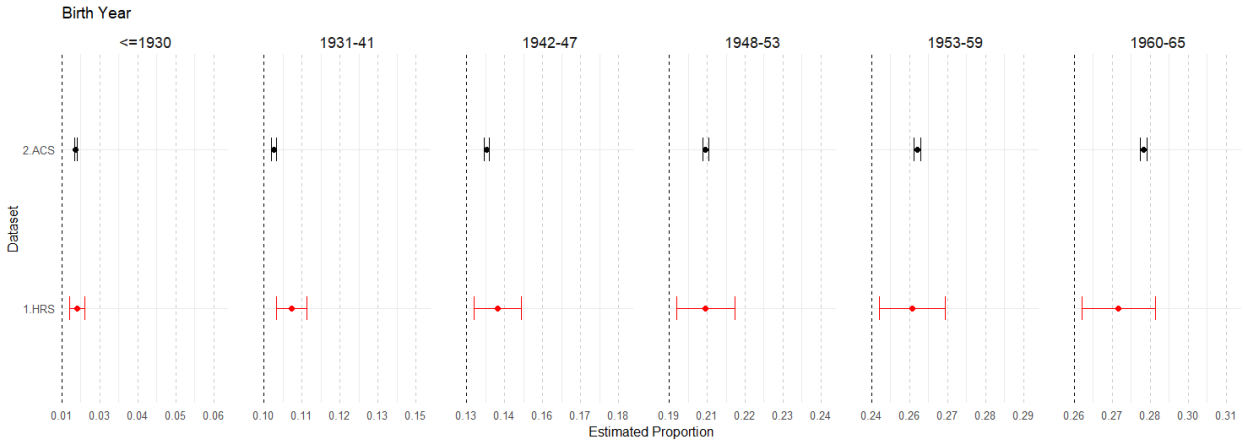
At the time of this writing, calibration adjustments are still being finalized for the 2022 HRS to incorporate individuals from the EGENX and MOC recruitment. Therefore, HRS estimates are subject to change upon our receipt of the final weights. Weighted estimates were calculated using the final survey weights provided in each dataset, with the HRS weights for panel

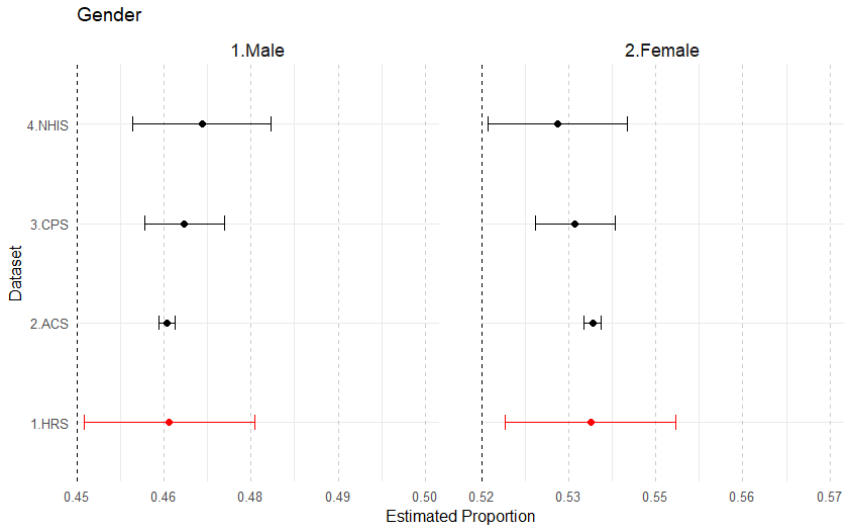
respondents considered “interim” or “early-release”. We also computed 95% confidence intervals for each descriptive parameter, taking the complex sampling features of each survey into account, and examined the overlap of these intervals across the data sources. The results are presented using error bar charts. Additional estimates and cumulative percentage plots are included in the available code. The table below summarizes the common variables analyzed.

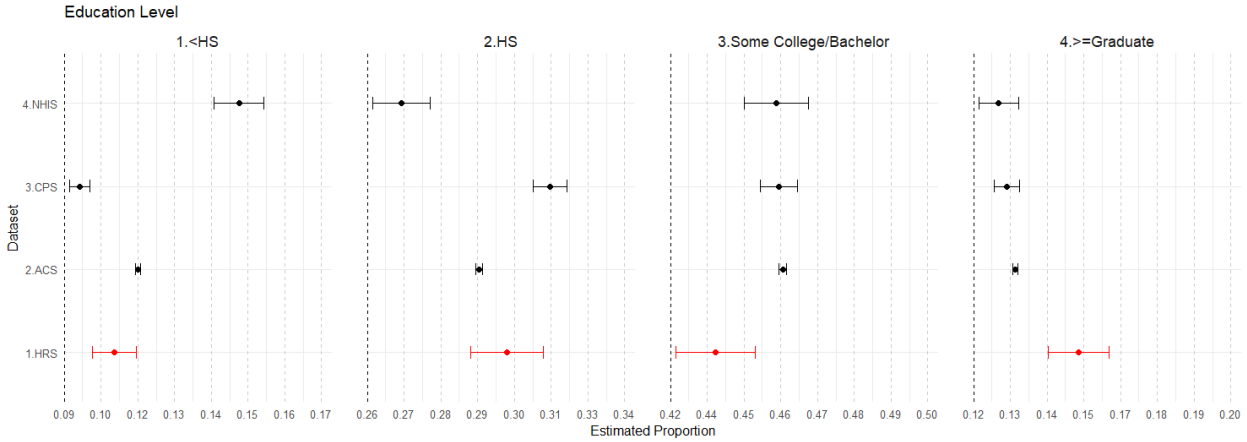
Variable	Variable type/Range	Available datasets
Age	Continuous	HRS; ACS; CPS; NHIS*
Birth Year	<=1930 1931-1941 1942-1947 1948-1953 1954-1959 1960-1965	HRS; ACS; CPS; NHIS**
Gender	Male Female	HRS; ACS; CPS; NHIS
Education Level	<=HS HS Some College/Bachelor's >=Graduate	HRS; ACS; CPS; NHIS
Race/Ethnicity	Hispanic Non-Hispanic Black Non-Hispanic Other	HRS; ACS; CPS; NHIS
Self-rated Health	Excellent Very Good Good Fair Poor	HRS; CPS; NHIS*
Self-Rated Health – Binary***	Good Not Good	HRS; CPS; NHIS*

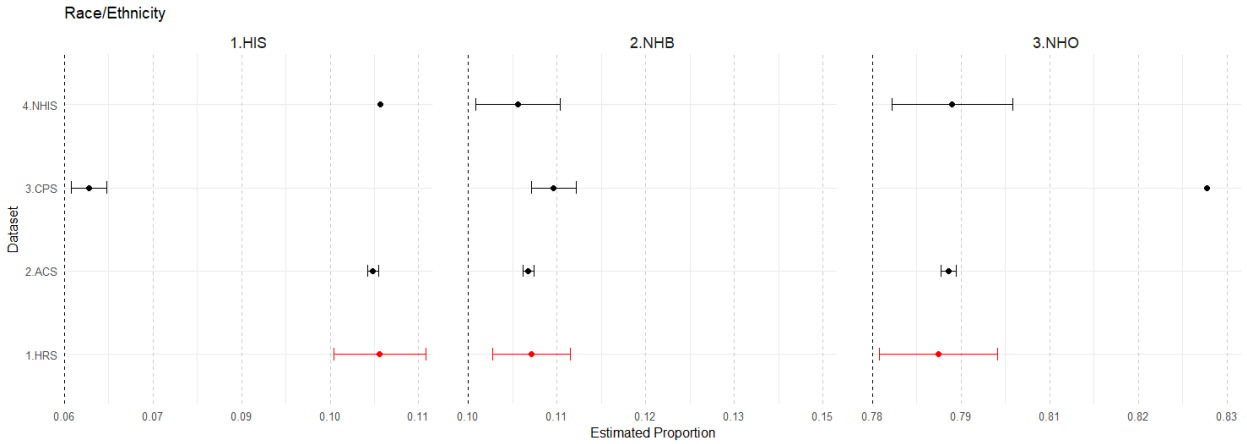
* CPS, age 80–84 was coded as 80, and age 85+ was coded as 85. In NHIS, age 85+ was coded as 85.
** CPS and NHIS do not provide birth year directly, so age cohorts were derived based on age at the time of the interview.
*** The binary self-rated health variable was recoded from the original categories as follows: "Excellent," "Very Good," and "Good" were grouped as "Good," while "Fair" and "Poor" were grouped as "Not Good."

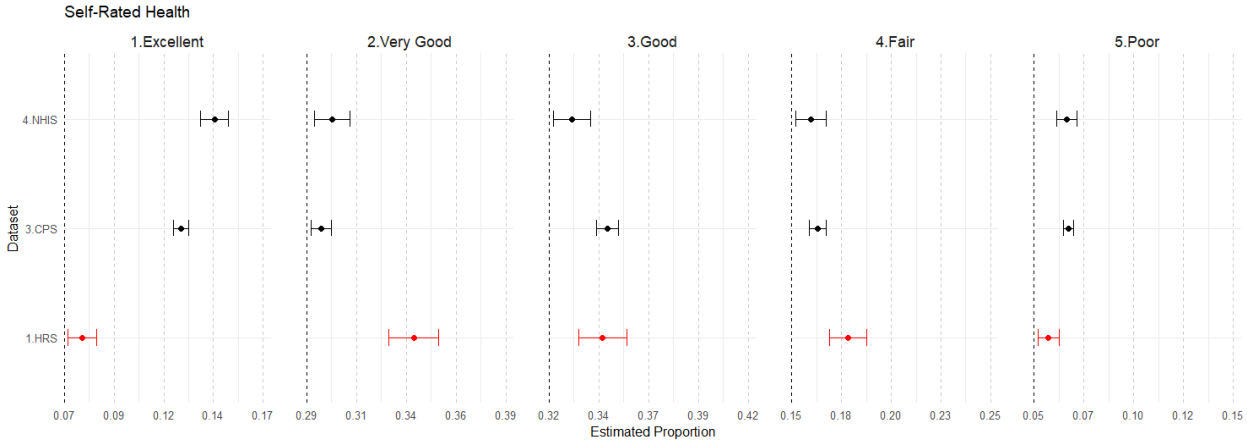

Age
4.NHIS
3.CPS
2.ACS
1.HRS
Dataset
65
66
67
68
69
70
Estimated Mean

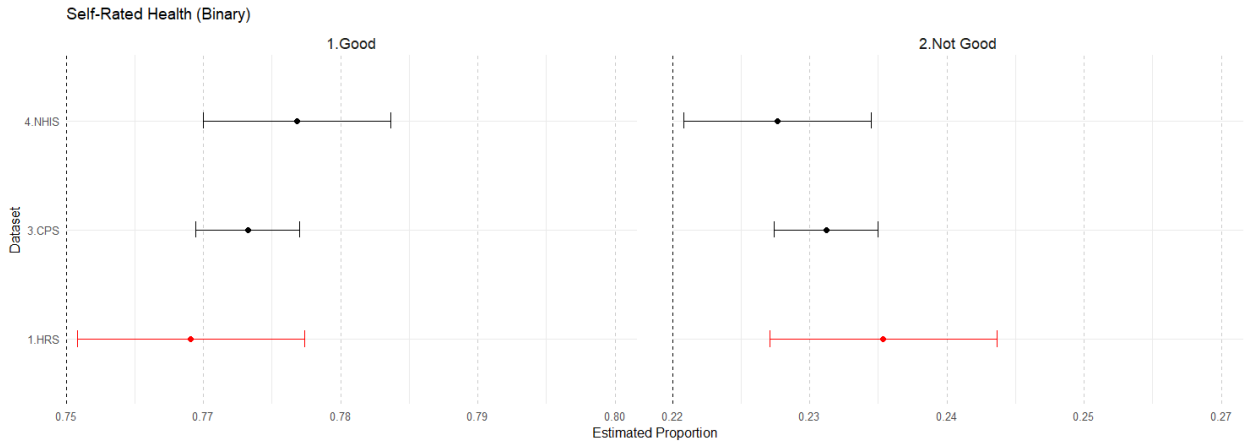

Birth Year
<=1930
1931-41
1942-47
1948-53
1953-59
1960-65
2.ACS
1.HRS
Dataset
0.01 0.03 0.04 0.05 0.06
0.10 0.11 0.12 0.13 0.15
0.13 0.14 0.16 0.17 0.18
0.19 0.21 0.22 0.23 0.24
0.24 0.26 0.27 0.28 0.29
0.26 0.27 0.28 0.30 0.31
Estimated Proportion


Gender
1.Male
2.Female
4.NHIS
3.CPS
2.ACS
1.HRS
Dataset
0.45 0.46 0.48 0.49 0.50
0.52 0.53 0.55 0.56 0.57
Estimated Proportion


Education Level
1.<HS
2.HS
3.Some College/Bachelor
4.>=Graduate
4.NHIS
3.CPS
2.ACS
1.HRS
Dataset
0.09 0.10 0.12 0.13 0.14 0.15 0.16 0.17
0.26 0.27 0.28 0.29 0.30 0.31 0.33 0.34
0.42 0.44 0.45 0.46 0.47 0.48 0.49 0.50
0.12 0.13 0.14 0.15 0.17 0.18 0.19 0.20
Estimated Proportion


Race/Ethnicity
1.HIS
2.NHB
3.NHO
4.NHIS
3.CPS
2.ACS
1.HRS
Dataset
0.06 0.07 0.09 0.10 0.11
0.10 0.11 0.12 0.13 0.15
0.78 0.79 0.81 0.82 0.83
Estimated Proportion


Self-Rated Health
1.Excellent
2.Very Good
3.Good
4.Fair
5.Poor
4.NHIS
3.CPS
1.HRS
Dataset
0.07 0.09 0.12 0.14 0.17
0.29 0.31 0.34 0.36 0.39
0.32 0.34 0.37 0.39 0.42
0.15 0.18 0.20 0.23 0.25
0.05 0.07 0.10 0.12 0.15
Estimated Proportion



Overall, outside of slight differences in the estimated proportions of the population in selected categories defined by education and self-rated health, the weighted HRS estimates seem well-aligned with the benchmark estimates from these other prominent national surveys. Additional calibration adjustments applied to the early-release HRS weights should align these estimates even further.

5. Fraction of Missing Information (FMI) for Key Variables

To calculate the Fraction of Missing Information (FMI) that is due to unit nonresponse, we imputed key HRS variables for 2022 non-responders using appropriate models for each key variable using information known before the start of data collection and the amount of effort expended during data collection. Our focus is on the FMI due to unit nonresponse, and therefore we exclude small numbers of cases with item-missing values for a given key variable or missing values on any predictor variables.

Step 1. Separately for each of the six key HRS variables considered in this analysis, we determine which predictors to include in an imputation model. This is accomplished by fitting a model using backward selection to identify the set of predictors that are significantly associated with the key variable among all cases that completed the interview in this wave (2022). The set of possible predictors considered included basic information from the tracker file and the amount of effort (in the form of in-person and telephone attempts) used during the 2022 data collection to get a 2022 interview. HRS tracker file variables include: Sex (male / female), Hispanic (yes / no), Spanish interview (yes / no), Currently married (yes / no), HRS Study cohort (HRS AHEAD CODA / WB / EBB / MBB / LBB), race (White / Black / Other), Education (<HS diploma or GED / HS diploma / some college / bachelor's degree / master's degree or more), Born in the US (yes / no), and birth year.

Step 2. We use the predictors found in Step 1 to impute missing values of each key variable for 2022 panel unit non-responders 200 times, using a stochastic multiple imputation approach that reflects imputation uncertainty.

Step 3. Calculate the FMI for each key variable and compare it to the simple unit nonresponse rate. Smaller FMIs indicate a greater amount of information being recovered from the multiple imputation process.

Based on the results summarized in the following table, not much information was recovered from imputation using the set of predictors we have available to us. For models with R-squared values above 0.2 (Work for Pay, Cognition Score, and Living Will) some information is gained, while models with less variation explained by our predictors (those with model R-squared values that are small), the FMI shows there is very little information gained from the imputation process. While these results suggest that these observable predictors cannot recover substantial amounts of the missing information in these variables, other observable variables not considered here may serve as stronger predictors, or the unit nonresponse may be non-ignorable (i.e., a function of the actual values on each key variable, even after conditioning on the observable predictors). The next section of the DQP considers this possibility.

Key Variable	Include proxy Rs?	n resp	n non-resp	Total possible n	R-sq from Step 1 model	Unit NR rate	FMI	Diff (Unit NR rate, FMI)
Work for Pay? (binary)	Yes	12,903	5690	18,593	0.2496	0.31	0.31	0.00
Cognition Score (0-27) (count)	No	11,663	5666	17,329	0.2100	0.33	0.27	0.06
CES-D symptom Cnt (0-8) (count)	No	12,186	5692	17,878	0.0671	0.32	0.33	-0.01
Self-rated Health (Excel/ VG) (binary)	Yes	12,922	5663	18,585	0.0779	0.30	0.32	-0.02
Nagi limitation Cnt (0-12) (count)	Yes	12892	5663	18555	0.1592	0.31	0.36	-0.05
Living will? (binary)	Yes	12695	5663	18358	0.2985	0.31	0.26	0.05

Analytic notes: Key variables that are counts are treated as continuous during the modeling and imputation process. Binary key variables were modeled using logistic regression.

6. Measures of Non-Ignorable Selection Bias

Using the weighted 2020 HRS core data as our population data source, we use the following set of covariates to calculate measures of non-ignorable selection bias for estimates related to five key HRS variables: education (HS diploma/GED or less, more than HS diploma), sex (male, female), ethnicity (Hispanic, non-Hispanic), race (black, non-black), and year of birth. The 2022 panel includes those 57 and older, and therefore we restrict the 2020 core data to those who are 57 and older.

Variable	Unweighted estimate (sample size)	Standardized Bias (95% credible interval)	Weighted estimate (95% confidence interval)	Adjusted estimate (95% credible interval)	Auxiliary Proxy Strength
Cognition	15.360 (11,830)	-0.120 (-0.283, -0.045)	16.014 (15.875, 16.153)	15.875 (15.554, 16.575)	0.400
Depression	1.482 (12,312)	0.298 (0.060, 1.428)	1.303 (1.247, 1.359)	0.888 (0, 1.362)	0.174
Self-rated Health (VG / Excellent) (binary)	0.352 (13,114)	-0.116 (-0.648, -0.036)	0.413 (0.397, 0.429)	0.468 (0.388, 1.000)	0.292
Nagi	3.817 (13,082)	0.183 (0.065, 0.485)	3.346 (3.237, 3.455)	3.182 (2.136, 3.593)	0.339
Work for pay (binary)	0.332 (13,038)	0.005 (-0.003, 0.024)	0.370 (0.354, 0.386)	0.327 (0.308, 0.335)	0.632

The table above presents the unweighted estimates and corresponding sample sizes, the standardized measure of unadjusted bias (SMUB; Little et al. 2020) estimates for continuous outcomes, and the measure of unadjusted bias for proportions (MUBP, Andridge et al. 2019) estimates for binary outcomes, each with 95% credible intervals, the weighted estimates (using the early-release survey weights for 2022), the model-based adjusted estimates with 95% credible intervals, and measures of auxiliary proxy strength. For continuous variables, the standardized bias column captures the median standardized departure of the unweighted mean from the population mean under a variety of possible non-ignorable nonresponse mechanisms. For binary variables, this column captures the median difference of the unweighted proportion from the population proportion under a variety of possible non-ignorable nonresponse mechanisms. Interpretation focuses on two features: whether the 95% credible interval excludes zero, indicating evidence of non-ignorable selection bias, and the magnitude of the estimate. Simulation studies have shown that when the auxiliary proxy is at least moderately predictive of the outcome, typically with correlations above 0.3 to 0.4, these outcome-aware bias indices reliably capture both the direction and magnitude of nonresponse bias (Martinez et al. 2026).

The estimated proportion of those who work for pay shows negligible bias (SMUB = 0.005; 95% CI: −0.003, 0.024) based on a strong proxy (0.632), suggesting that nonresponse for this outcome is either largely explained by the observed covariates, or occurring entirely at random (consistent with the earlier FMI analysis). The estimated mean cognition score and the estimated proportion with excellent or very good self-rated health show small negative biases (SMUB = −0.120 and SMUB = −0.116, respectively), with credible intervals that exclude zero, indicating that respondents tend to score lower on both outcomes than nonrespondents even after covariate adjustment. The mean depression score yields a relatively small SMUB of 0.298 with a wide credible interval (0.060, 1.428), reflecting substantial uncertainty that is likely driven in part by the weak proxy strength for this outcome (0.174). It is worth noting that a complete assessment of nonresponse bias for depression would require outcome-specific predictors with stronger associations, which is outside the scope of this analysis, since we are using the same set of predictors for all key variables. The estimated mean of Nagi functional limitations also

shows evidence of a small non-zero bias (SMUB = 0.183; 95% CI: 0.065, 0.485), with moderate proxy strength (0.339).

The early-release weights shift the estimates in the apparently correct direction across all outcomes, consistent with the model-based adjusted estimates. The exception is work for pay, where the weights appear to over-adjust, as the weighted estimate (0.370) moves away from (rather than toward) the model-based adjusted estimate (0.327). Overall, these results present negligible evidence of substantial non-ignorable selection bias in these estimates and, therefore, limited evidence of nonresponse bias beyond what can be addressed through traditional survey weighting. We note that these findings remain preliminary and will be updated once the final 2022 HRS survey weights become available.

7. Alternative Weighted Estimates

This analysis requires access to finalized 2022 weights for all panel cases. Post-stratification of 2022 base weights is ongoing at the time of this writing, and therefore this analysis is on hold until that process is complete. For our set of five key variables, we will eventually compare unweighted estimates, base-weighted (interim) estimates, and post-stratified (final) estimates of means; see the table below. The purpose of this analysis is to quantify the impacts of the various weighting steps on the point estimates and the standard errors.

Variable	Estimated mean based on interim 2022 weight	Standard Error	Estimated mean based on final 2022 weight	Standard Error
Cognition	16.014	0.070	TBD	TBD
Depression	1.303	0.029	TBD	TBD
Self-rated Health (VG/Excellent)	0.413	0.008	TBD	TBD
Nagi	3.346	0.057	TBD	TBD
Work for pay	0.370	0.008	TBD	TBD

8. Summary of Nonresponse Follow Up Strategies

The HRS utilizes interviewers to invite, encourage, and facilitate interviews with its panel members each wave. The 2022 HRS sample was divided into three groups with slightly different protocols for each group. Group 1 are those assigned to in-person data collection, Group 2 are those assigned to telephone, and Group 3 are those assigned to a web-first protocol. Groups 1 and 2 are sent initial invitation letters, at which point interviewers begin making contact attempts to get an interview, either via the telephone or in-person. Nonresponse experiments for those in Groups 1 and 2 who do not initially complete the interview are discussed in Section 9 of this DQP. The nonresponse protocol for those in Group 3 who do not complete the web survey

during the initial 4-week period (which includes a mailed/emailed invitation and several mailed/emailed reminders) includes interviewers beginning to make telephone calls to schedule a telephone interview as a form of nonresponse follow-up.

9. Summary of Other Qualitative Strategies

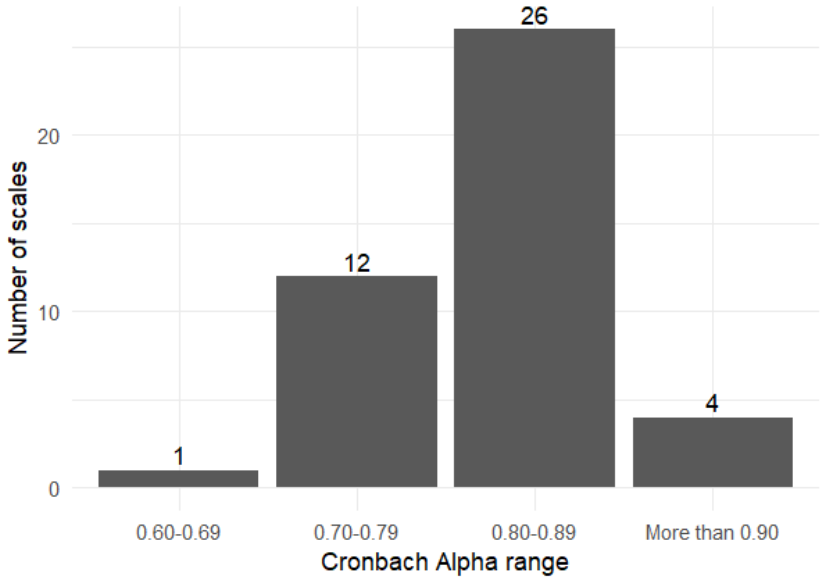
To improve sample balance and representativeness, the HRS engages in several responsive survey design strategies that are implemented in real-time during data collection. These include **case prioritization**, where active sample cases predicted to have a meaningful influence on multiple key survey estimates are prioritized with interviewer effort; **end game strategies**, where active sample cases that have reached a certain effort threshold without responding are provided with a one-time offer of only one additional face-to-face attempt, with an increased incentive offer and possibly an option to complete the survey by web as well; **mode switching**, where ideal modes according to the survey design may not be feasible / desired for a given panel member and alternative modes are employed instead to collect the data (e.g., face-to-face to telephone); and **limited effort / stopping rules**, where decisions are made to discontinue effort on difficult cases that are hard refusals and/or unlikely to shift key estimates in a meaningful fashion. The data collection team also regularly analyzes paradata describing interviewer effort and the results of contact attempts, in an attempt to tailor interviewer efforts and maximize the efficiency of interviewer contact attempts.

Measurement Error

10. Reliability: Cronbach's Alpha

Sections 10 and 11 of this report consider standard approaches to assessing measurement quality. We focus on 43 scales from the HRS Leave-Behind Questionnaire that were designed as psychometric measures, administered in at least three waves, and composed of three or more items for this analysis. To first assess internal consistency (as one possible measure of reliability), we calculated Cronbach's alpha separately for each scale using its corresponding item battery. As shown in the figure below, the distribution of alpha coefficients indicates that most scales exhibit moderate to high internal consistency. In particular, 26 scales fall in the 0.80–0.89 range, corresponding to excellent reliability. Another 12 scales fall in the 0.70–0.79 range, which is generally considered as reflecting acceptable internal consistency. Four scales exceed 0.90, indicating very high internal consistency, while only one scale falls below 0.70. Overall, the majority of the HRS Leave-Behind scales meet conventional reliability standards for psychometric analysis.

Additional details on scale construction, item wording, and scoring procedures can be found in the Psychosocial and Lifestyle Questionnaire 2006–2022 User Guide.



Distribution of Cronbach's alpha coefficients for the 43 psychosocial scales from the HRS Leave-Behind Questionnaire. Bars show the number of scales falling within each range.

11. Confirmatory Factor Analysis

Based on input from HRS co-Is and other researchers working with HRS data, we evaluated the dimensionality of seven scales found in either the core HRS interview or the leave behind questionnaire using confirmatory factor analysis (CFA). We did this by specifying a single latent factor for each construct suggested from this input. This analysis was conducted to assess whether the items within each scale adequately reflected a unidimensional underlying construct and demonstrated acceptable model fit.

Models were primarily estimated using the Weighted Least Squares Mean and Variance (WLSMV) estimator to account for the ordinal nature of the survey items. For the physical demand scale from the main interview (Section J), one-factor models were estimated separately by survey mode (face-to-face, telephone, and web) to assess potential mode-specific differences in measurement fit. All other scales are in the leave-behind paper questionnaire, which is given to respondents after they complete the in-person survey. Maximum likelihood (ML) estimation was used for scales using continuous items.

Scale	Mode	CFI	TLI	RMSEA	SRMR	χ^2
WLSMV						
Physical demand	Face-to-Face	0.915	0.830	0.327	0.162	401.15
	Telephone	0.954	0.908	0.227	0.127	209.65
	Web	0.947	0.893	0.409	0.131	152.15

Religiosity	Paper	0.999	0.998	0.142	0.020	152.02
Anxiety symptoms		0.993	0.987	0.091	0.045	161.06
Life satisfaction		0.999	0.996	0.120	0.025	108.39
ML						
Religiosity	Paper	0.971	0.912	0.232	0.023	405.16
Anxiety symptoms		0.935	0.869	0.149	0.047	418.98
Life satisfaction		0.969	0.906	0.198	0.028	293.40

Note. WLSMV was used when indicators were treated as ordinal, and ML was used when indicators were treated as continuous. Both estimators were considered as sensitivity checks.
Items for CFA model. Physical demand: physical effort; lifting heavy loads; stooping, kneeling, or crouching; good eyesight; intense concentration. Religiosity: belief in God; events unfold according to a divine plan; trying hard to carry out religious beliefs; finding strength in religion. Anxiety symptoms: fear of the worst happening; nervousness; trembling hands; fear of dying; feeling faint. Life satisfaction: life is close to ideal; life conditions are excellent; satisfied with life; have gotten the important things in life.

Model fit was evaluated using the comparative fit index (CFI), Tucker–Lewis index (TLI), root mean square error of approximation (RMSEA), standardized root mean square residual (SRMR), and the chi-square statistic. Following common guidelines, CFI and TLI values above 0.90 and RMSEA and SRMR values below 0.08 were considered indicative of acceptable fit. Because of the large sample size, chi-square statistics were reported for completeness but were not used as the primary criterion for model evaluation.

Across constructs, CFI and TLI values generally exceeded conventional thresholds under WLSMV estimation, suggesting that the items load consistently on a single latent factor. However, higher RSMEA and SRMR values raised concerns about the quality of the scale fits. RMSEA values were elevated for several models, particularly for the physical demand scale across interview modes and for the leave-behind scales under ML estimation. SRMR values were generally low across the models, suggesting that the differences between the observed item relationships and the relationships implied by the model were fairly small.

Overall, these results provide only partial support for a one-factor structure. The physical demand scale showed relatively poor fit across all interview modes, suggesting weaker unidimensionality. The leave-behind scales generally fit better under WLSMV than under ML, although some evidence of misfit remained. Taken together, the findings suggest that the scales are broadly consistent with a single underlying construct, but the fit is not uniformly strong across all measures.

12. Interviewer Variance in Key Variables

The table below presents estimated interviewer variance components and intraclass correlation coefficients for selected key HRS variables measured among 2022 panel cases. These estimates are based on simple multilevel models and do not include adjustments for any other covariates, providing a “raw” sense of the interviewer variance in each measure. We also note that these estimates assume that each interviewer is working a random sample of the full HRS sample, which is not the case given the HRS sample design. The estimates provide an initial sense of the between-interviewer variance, which may be explained by area- and respondent-level covariates.

Variable	Estimated Interviewer variance	Likelihood Ratio Test	Intraclass Correlation Coefficient
Cognition	0.50	136.4***	2.7%
Depression	0.03	33.9***	0.9%
Self-rated Health	0.05	103.6***	1.5%
Nagi	0.17	88.5***	1.4%
Work for pay	0.19	84.7***	3.0%

The analysis included 13,111 cases nested within 253 unique interviewers (about 52 cases per interviewer on average). Eight interviewers who appeared only once were excluded. Depending on the outcome, we fitted linear or logistic mixed-effects models with an interviewer-specific random intercept and compared each model to a corresponding null model without interviewer effects.

The above table reports the estimated interviewer-level variance, likelihood ratio tests, and intraclass correlation coefficients (ICCs) for the five key variables. Across all outcomes, the likelihood ratio tests strongly reject the null hypothesis of no interviewer-level variance ($p < 0.001$), indicating statistically significant interviewer effects. However, the magnitudes of the interviewer-level clustering varied across measures. Work for pay showed the largest ICC (3.0%), followed by cognition (2.7%), self-rated health (1.5%), and Nagi functional limitations (1.4%). Depression (0.8%) exhibited smaller but non-negligible interviewer-level variation.

Although these ICCs are relatively small in magnitude, they consistently indicate measurable clustering of responses at the interviewer level. This clustering may be explained by other area or respondent-level covariates. The fact that these are “raw” estimates of ICCs prior to any covariate adjustment suggests that interviewer effects are not a significant problem in the measurement of these variables.

13. Straightlining and Speeding

Following Leiner (2019), we constructed an indicator of unusually fast section completion using section-level response times from five sections of the 2022 instrument. The analysis is restricted to respondents who completed the first five sections (A–E) of the 2022 main HRS interview either face-to-face or by telephone and for whom section-level response times were recorded (n = 11,655).

For each section, we calculated a speed factor as the ratio of the median section completion time to an individual's completion time, capped the value at three, and averaged the capped speed factors across the five sections. A higher speed factor indicates faster-than-typical completion relative to other respondents. For descriptive purposes, we further classified the continuous speed factor index into four categories based on their speed factor: slower than typical, slightly faster than typical, moderately faster than typical, and very fast.

The table below presents the distribution of the section-level speed factor index overall and by interview mode. Only 0.2% of responders had an average speed factor of 2 or higher. Differences by interview mode were modest. In both modes, extremely rapid completion was uncommon, indicating that unusually fast responding was not strongly associated with interview mode.

Distribution of the Section-Level Speed Factor Index

Classification	**Criterion**	**Mode**		
		Face-to-face (n=5595)	**Telephone (n=5402)**	**Total**
Slower than typical (median)	0≤speed factor≤1	44%	45%	45%
Slightly faster	1<speed factor≤1.5	45%	43%	44%
Moderately faster	1.5<speed factor≤2	10%	9%	10%
Very fast	2<speed factor≤3	0.1%	0.3%	0.2%

Distribution of Straightlining Factor by Questionnaire Section and Data Collection Mode

Section		**Mode**			
		Face-to-Face	**Telephone**	**Web**	**Total**
J (Employment, 14 Likert-type items) (n=)	Strict	0.1%	0.0%	0.4%	0.11%
	Moderate	0.8%	0.6%	0.6%	0.6%

P (Expectations, 14 items, 0-100 scale)	Strict	6.0%	4.5%	3.2%	5.0%
	Moderate	8.5%	7.5%	5.3%	7.6%

Straightlining was assessed separately within two sections of the questionnaire: section J and P. We calculated the within-section standard deviation (SD) of responses across all items in each section. This approach provides a scale-agnostic measure of response differentiation and allows consistent assessment across scales that differ in item format (Kim et al., 2019; McCarty & Shrum, 2000). A straightlining indicator was defined as "strict" when the within-section SD = 0, indicating identical responses across all items, and as moderate when the SD was greater than zero but less than or equal to 0.25, reflecting limited response variation within the section. Overall, the results in the table above present minimal evidence of straightlining behavior under both definitions.

14. Breakoff Rate

To define a "breakoff" in the HRS panel, we first need to know the definition of an accepted partial interview. The definition of an accepted partial interview for respondents (which does not apply to exit or post-exit interviews) to the core HRS interview is completing the delayed word recall test in the cognition section (section D). Interviews suspended prior to the delayed word recall test in section D are considered "breakoffs".

Per the table below, the breakoff rate among panelists who started the 2022 HRS core survey is 0.6%. This estimate includes all responses from all three modes; in-person, telephone and web. We did not attempt to break these rates down further by mode at this time because the overall rate is low and does not cause concern. Further, the mode used to start a survey is not readily available for non-complete cases.

Interview status	Count (% of started)
Started	13,300
Completed	13,087 (98.4%)
Accepted partial	134 (1.0%)
Break off (before accepted partial cutoff)	79 (0.6%)

*Note: counts in this table are produced before data finalization takes place. Counts in this table may not match counts from finalized documentation.

15. Mode Effects on Key Variables

To evaluate potential mode effects, we compared responses on five key variables across face-to-face, telephone, and web interviews. The table below presents the estimated mean (or proportion) of each key variable by interview mode. One-way ANOVA was used for continuous outcomes, and chi-square tests were used for categorical outcomes.

Unweighted Mode effects for five Key Variables by Interview Mode

Variable	FTF	Telephone	Web	Total	Test and Effect Size
Cognition	14.988 (n=5,485)	15.986 (n=4,907)	14.643 (n=1,438)	15.360 (n= 11,830)	F statistic: 9.943*** η2: 0.01
Depression	1.487 (n=5,781)	1.563 (n=5,162)	1.156 (n=1,369)	1.482 (n=12,312)	F statistic: 10.16*** η2: <0.1
Self-rated Health	0.345 (n=5,880)	0.322 (n=5,767)	0.485 (n=1,467)	0.352 (n=13,114)	X-squared: 136.16*** Cramer's V: 0.10
Nagi	3.933 (n=5,872)	3.943 (n=5,746)	2.740 (n=1,464)	3.817 (n=13,082)	F statistic: 62.3*** η2: <0.1
Work for pay	0.306 (n=5,851)	0.358 (n=5,732)	0.334 (n=1,455)	0.332 (n=13,038)	X-squared: 35.43*** Cramer's V: 0.05

Note: F tests are from one-way ANOVA for continuous outcomes. Chi-square tests are used for categorical outcomes. η^2 is eta squared. V is Cramér's V. *** $p < 0.001$.

All five variables differed statistically across modes ($p < 0.001$). However, the estimated effect sizes were uniformly small (η^2 at or below 0.01 for continuous outcomes and Cramér's V below 0.10 for categorical outcomes), indicating that although differences are statistically detectable given the large sample size, their practical magnitude of the differences is modest at best.

Overall, telephone respondents had higher cognition scores, higher depression scores, and a higher proportion of working for pay. Web respondents reported lower depression and more frequently reported excellent or very good self-rated health. Despite these directional patterns, the small effect sizes suggest a limited impact of interview mode on these key variables.

These results are unweighted because the web subsample was not selected to support population inference. The purpose of this analysis is instead to examine potential mode effects by comparing observed responses across interview modes, rather than to generate weighted population estimates.

16. Summary of Survey Design Steps

A summary of the initial 1992 HRS design can be found here.

The HRS survey content was created by four expert working groups, including Labor Force Participation and Pensions, Health Conditions and Health Status, Family Structure and Mobility, and Economic Status. The HRS co-Is continually monitor the HRS core content to maintain the relevance of the HRS data. They incorporate feedback from both the HRS data user community and feedback received from HRS interviewers about respondent experiences. Each wave HRS conducts a pre-test with a small subset of HRS respondents to rigorously test, identify, and address survey instrument concerns. Data collected from this set of pre-test respondents is not included in final HRS data products.

As the HRS panel population ages, some respondents become unable to directly participate, and therefore a proxy respondent must be identified. Careful work is done by interviewers and project staff to select proxies to help respondents continue to participate as they begin to experience physical and mental health limitations. Further details on proxy selection can be found here.

17. Summary of Editing and Imputation Steps

Most sections of the core HRS data released to the public are not imputed. The one section that has a regular imputation process is the cognition section. Details of the cognition imputation work can be found here. Select variables found in RAND data products that are created from core HRS data have imputed variables. More information on RAND imputations can be found here.

Sampling Error

18. Target Population

The target population for the HRS is the US population aged 51 and older who live in the contiguous United States who reside in a household, and their younger spouses. Following conventional practice for population surveys, institutionalized persons (those in prisons, jails, nursing homes, and long-term or dependent care facilities) are excluded from the survey population. The original sample and subsequent new cohorts recruited into the HRS exclude institutionalized persons. Panelists who subsequently move into nursing home facilities as they age remain in the sample. Therefore, we have nursing home representation in the HRS panel. For more details, see the Original Sample Documentation, or the HRS 2016 Weighting Documentation.

19. Achieved Sample Size

The achieved sample size for the HRS 2022 panel was 13,128 accepted complete interviews. (RR website).

20. Design Effects and Coefficients of Variation for Key Variables

Variable	Estimate	Standard Error	Design Effect (DEFF)	Coefficient of Variation (CV)
Cognition	16.014	0.070	3.182	0.44%
Depression	1.297	0.029	2.756	2.25%
Self-rated Health (Excellent / VG)	0.413	0.008	3.592	2.06%
Nagi	3.347	0.057	3.634	1.69%
Work for pay	0.367	0.008	3.533	2.25%

As shown in the above table, the design effects range from approximately 2.8 to 3.6, implying that the variances are about three to four times larger than they would be under simple random sampling. This reflects the influence of the complex sample design, including stratified cluster sampling and unequal weighting. In contrast, coefficients of variation are low for all estimates, remaining below 2.5 percent, indicating good precision despite the variance inflation.

21. 1 + L / Unequal Weight Effect

The overall unequal weighting effect (Kish's 1+L value) for the 2022 HRS panel is 2.03, indicating that variability in the final analysis weights inflates the variance of weighted estimates by approximately 103 percent relative to an equal probability sample of the same size. This quantity represents a worst-case scenario for variance inflation due to weighting alone. It serves as an upper bound under standard assumptions of stratified sampling and approximate independence between the outcome variable and the survey weights, abstracting from additional design features such as clustering.

22. Brief Sampling Descriptions

When recruiting new cohorts, the HRS selects a stratified multistage cluster probability sample of primary sampling units (U.S. counties or groups of counties), area segments, and households, with oversampling of households likely to have persons from the target new cohort present (based on linked commercial data). Households initially are invited to complete a screening interview, followed by the main baseline interview. Subsequent interviews with panel members who completed the baseline survey are attempted either face-to-face or by telephone with alternating half-samples of active panel members, with web also offered as an option for those randomly assigned to the telephone mode. Additional details on the sample design can be found here.

23. Link to Analytic Guidelines

Analytic guidelines for users of the HRS data, including examples of syntax for statistical software packages, can be found here: https://hrs.isr.umich.edu/documentation/new-user-guide.

References

Andridge R. R., West B.T., Little R. J. A., Boonstra P. S., and Alvarado-Leiton F. (2019). "Indices of Non-Ignorable Selection Bias for Proportions Estimated from Non-Probability Samples". In: Journal of the Royal Statistical Society Series C: Applied Statistics 68.5, pp. 1465–1483. ISSN: 0035-9254. DOI: 10.1111/rssc.12371. URL: https://doi.org/10.1111/rssc.12371

Kim, Y., Dykema, J., Stevenson, J., Black, P., & Moberg, D. P. (2019). "Straightlining: Overview of measurement, comparison of indicators, and effects in mail–web mixed-mode surveys". Social Science Computer Review, 37(2), 214-233.

Leiner, D. J. (2019, December). "Too fast, too straight, too weird: Non-reactive indicators for meaningless data in internet surveys". In Survey Research Methods (Vol. 13, No. 3, pp. 229-248).

Little R. J. A., West B.T., Boonstra P.S., and Hu J. (2020). "Measures of the Degree of Departure from Ignorable Sample Selection". In: Journal of Survey Statistics and Methodology 8.5, pp. 932964. ISSN: 2325-0984. DOI: 10.1093/jssam/smz023. URL: https://doi.org/10.1093/jssam/ smz023.

Martinez S.D., West B.T., Andridge R.R. (2026) Measures of non-ignorable selection bias for non-probability samples. The Survey Statistician, 93, 26-41. (Editor Reviewed)

McCarty, J. A., & Shrum, L. J. (2000). "The measurement of personal values in survey research: A test of alternative rating procedures". Public Opinion Quarterly, 64(3), 271-298.

Schouten, B., Cobben, F., & Bethlehem, J. (2009). Indicators for the representativeness of survey response. *Survey Methodology*, *35*(1), 101-113.