INSTITUTE FOR SOCIAL RESEARCH
SURVEY RESEARCH CENTER
UNIVERSITY OF MICHIGAN

A Complete Example of the SRC Data Quality Profile from the 2022 Wave of the Health and Retirement Study (HRS)

Brady T. West and Heather M. Schroeder

bwest@umich.edu, schroedh@umich.edu

Acknowledgements

This was certainly a team effort, and we are grateful to the HRS leadership and staff for their consistent support and review.

We are especially grateful to **Chendi Zhao** and **Sergio Martinez**, each PhD candidates in PSDS, who have contributed in countless ways to the methodological work of the HRS and played a tremendous role in completing this data quality profile.

Key variables used throughout the DQP

Lesson 1: A project needs to have a thorough discussion about "key" variables in the eyes of the project (important to stakeholders, widely used in publications, etc.). These may also be measures of change!

Selected HRS Key Variables:

- Work for Pay (binary)
  - Are you currently working for pay? (yes/no)
- Cognition Score (0-27) (treated as continuous)
  - Summary of 3 cognition tests: immediate word recall (max of 10), delayed word recall (max of 10), subtract 7 from 100 (5 times), count backwards from 20
- CES-D Score (0-8) (treated as continuous)
  - Count of "yes" responses to 8 CES-D depression symptom questions
- Excellent or very good self-rated health (binary)
  - Original response categories: Excellent, very good, good, fair, poor
- Nagi limitation count (0-12) (treated as continuous)
  - Count of "yes" responses to 12 questions about difficulties with daily living activities

Reminders: The SRC DQP

The purpose of the Data Quality Profile (DQP) is to objectively describe the quality of the HRS 2022 panel survey from the Total Survey Error (TSE) perspective.

The DQP 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**:

- Nonresponse error / selection bias
- Measurement error
- Sampling error

Additional HRS Context / Details

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.**

Domain 1:
Nonresponse Error / Selection Bias

1. AAPOR (RR6) Response Rate

The AAPOR RR6 response rate (deemed most appropriate for the HRS design) was **67.9%** among 2022 panelists.

Reminder: Does this provide evidence of nonresponse bias?

Lesson 2: What AAPOR response rate makes the most sense for your project?

2. Overall R-Indicator

The overall R-indicator, as a measure of the representativeness of the realized sample based on variability in predicted response propensities across subgroups, was **0.682** (1 = perfect, 0 = lowest possible).

Lesson 3: BART was the best-fitting model for estimating response propensities (**AUC = 0.793**) among several machine learning approaches, using household-level predictors from the 2020 HRS.

This result suggests that there is variation in response propensity among relevant subgroups based on observable variables, and if these subgroups also vary in terms of values on key variables, nonresponse adjustment will be needed.

2. Partial R-Indicators

The unconditional partial R-indicator (as a measure of how much of the variation in response propensities is due to a GIVEN variable) for **COHORT** (crucial for HRS data users) was **0.0258**.

This measure is capped at 0.5 (as a maximum contribution), so COHORT explains only a small fraction of this variability. COHORT plays only a minor role.

The category-level partial R-indicators for the categories of COHORT can be used to identify over- or under-representation:

Category	Partial R-Indicator
Oldest cohorts (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

This could be repeated for other key predictors of response propensity. LBBs (RR = 59.7%) are relatively under- represented in the realized sample.

3. Subgroup Variation in Response Propensity

Non-Hispanic Whites/Others and more 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, but this approach assumes that nonresponse was not a function of other key variables that could not be observed from non-respondents.

AAPOR RR6 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%

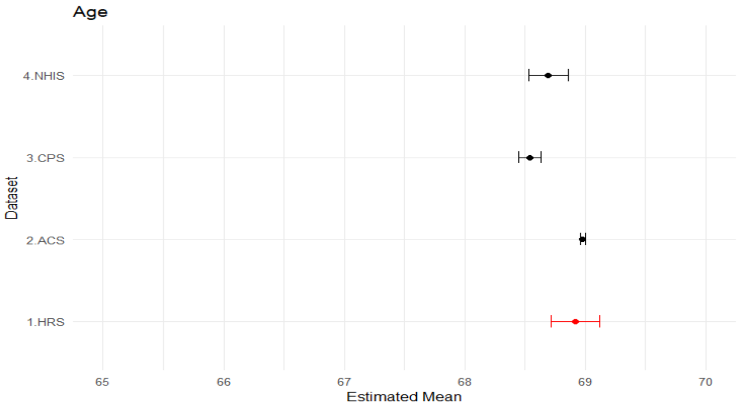
4. References to Benchmarks

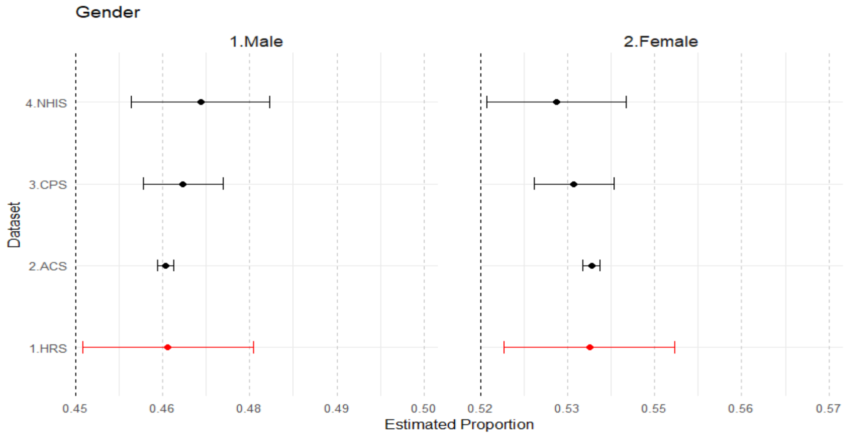
Important Note: At the time of this writing, calibration adjustments are still being finalized for the 2022 HRS to incorporate individuals from the EGENX and MOC recruitment. **Therefore, all HRS estimates in this report are subject to change upon our receipt of the final weights.** The HRS weights for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

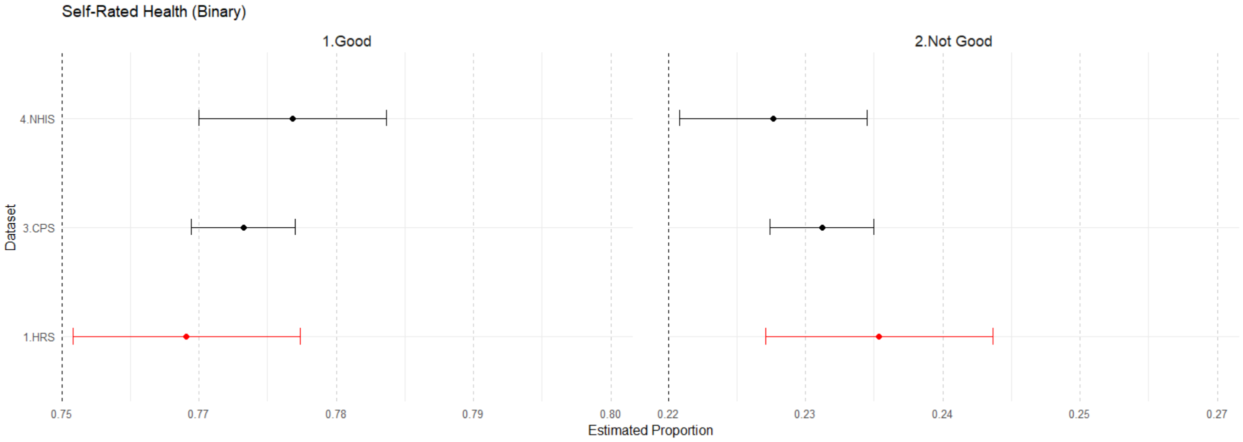
Using the early-release weights and focusing on those 57+ (the age of the youngest panel member), we **compared weighted HRS estimates** (and 95% confidence intervals) for mean age, birth year, sex, education, race/ethnicity, self-rated health, and binary self-rated health (good vs. other) from the **2022 ACS, the 2022 CPS, and the 2022 NHIS**.

4. References to Benchmarks

Overall, prior to the calibration adjustments, **the HRS estimates are largely in-line with estimates from these other major national surveys.**







5. Fraction of Missing Information (FMI)

The FMI provides a measure of how much information can be recovered from existing auxiliary variables when imputing missing data.

We followed three steps:

1. Determine what 2020 predictors to use to impute **unit nonresponse** for each of our key variables.
2. Stochastically impute each missing value 200 times.
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.

5. Fraction of Missing Information (FMI)

Overall, we could recover some information for cognition score and the living will indicator, but for other variables, the nonresponse may be occurring completely at random, or could be non-ignorable.

Key Variable	Include proxy Rs?	Resps	N	R-sq: Step 1 model	Unit NR rate	FMI	Diff (Unit NR rate, FMI)
Work for Pay? (binary)	Yes	12,903	18,593	0.2496	0.31	0.31	0.00
Cognition Score (count)	No	11,663	17,329	0.2100	**0.33**	**0.27**	**0.06**
CES-D symptoms (count)	No	12,186	17,878	0.0671	0.32	0.33	-0.01
Self-rated Health (binary)	Yes	12,922	18,585	0.0779	0.30	0.32	-0.02
Nagi limitations (count)	Yes	12892	18555	0.1592	0.31	0.36	-0.05
Living will? (binary)	Yes	12695	18358	0.2985	**0.31**	**0.26**	**0.05**

6. Measures of Non-Ignorable Selection Bias

If nonresponse is not occurring at random, and is a function of the variables of interest even after conditioning on the auxiliary information, weighting will not entirely correct the bias.

Using the 2020 HRS core data as a population source and identifying a unique set of scientifically relevant predictors of each key variable for the 2022 panel respondents that were also measured in 2020, we computed:

1. The SMUB measure (and an adjusted estimate based on the SMUB) for the means of the continuous variables (Little et al., 2020), and
2. The MUBP measure (and an adjusted estimate) for the proportions based on the binary variables (Andridge et al., 2020)

We also computed weighted estimates using the early-release weights.

6. Measures of Non-Ignorable Selection Bias

Overall, **there was minimal evidence of non-ignorable selection bias**, and at least some of the potential selection bias was being repaired (at least in part) by the early-release weights. Non-ignorable selection bias does not seem to be an issue for these key estimates.

Key Variable	Unweighted estimate (sample size)	Standardized Bias (95% credible interval)	Weighted estimate (95% confidence interval)	Adjusted estimate (95% credible interval)	Auxiliary Proxy Strength
Cognition Score	15.36 (11,830)	-0.12 (-0.28, -0.05)	16.01 (15.88, 16.15)	15.88 (15.55, 16.58)	0.40
CES-D symptoms	1.48 (12,312)	0.30 (0.06, 1.43)	1.30 (1.24, 1.36)	0.89 (0, 1.36)	0.17
Self-rated Health	0.35 (13,114)	-0.12 (-0.65, -0.04)	0.41 (0.40, 0.43)	0.47 (0.39, 1.00)	0.29
Nagi limitations	3.82 (13,082)	0.18 (0.07, 0.49)	3.35 (3.24, 3.46)	3.18 (2.14, 3.59)	0.34
Work for pay?	0.33 (13,038)	0.01 (-0.003, 0.02)	0.37 (0.35, 0.39)	0.33 (0.31, 0.34)	0.63

7. Alternative Weighted Estimates

We could not consider alternative weighted estimates, given that the calibration adjustments are still being computed for the overall 2022 core weights.

In other applications, one would look at shifts in the estimates (and their standard errors) from each step of the weighting process, to assess the impacts of the weighting adjustments on inference.

8. Summary of NRFU 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

- Group 1 are those assigned to in-person data collection
- Group 2 are those assigned to telephone
- 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 follow-up strategies for those in Groups 1 and 2 who do not initially complete the interview are discussed in **Section 9** (to follow).

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. These include:

1. **Case Prioritization**, where active sample cases predicted to have a meaningful influence on multiple key survey estimates are prioritized with interviewer effort;

2. **Endgame 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;

3. **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

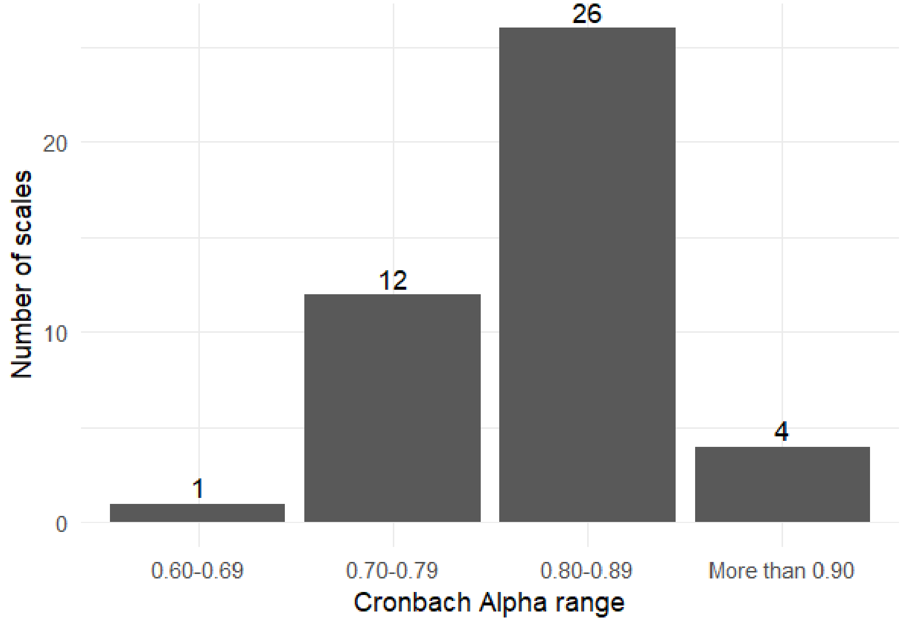
4. **Limited Effort / Stopping Rules**, where decisions are made to discontinue effort on difficult cases that are hard refusals and/or unlikely to shift key estimates in a meaningful fashion. The data collection team also regularly analyzes paradata describing interviewer effort and the results of contact attempts, to tailor efforts.

Domain 2:
Measurement Error

10. Reliability: Cronbach's Alpha

We assess the internal consistency of 43 scales from the HRS Leave-Behind Questionnaire.

- Calculate Cronbach's alpha for each scale using it's battery of items
  - High reliability: 0.90 +
  - Excellent reliability: 0.80 – 0.89
  - Acceptable reliability: 0.70 - 0.79
  - Lower reliability: <0.70



Overall, the majority of the HRS Leave-Behind scales have items that meet conventional standards for high reliability.

INSTITUTE FOR SOCIAL RESEARCH
SURVEY RESEARCH CENTER
UNIVERSITY OF MICHIGAN

11. Confirmatory Factor Analysis

We also performed CFA for several selected scales to assess whether the items reflect a unidimensional underlying construct. **Concerns arise with the physical demand scale in particular**, which suggests caution with analyzing this as a single scale (WLSMV is best for categorical items).

		Model Fit Statistics			
Scale	**Mode**	**CFI (>0.9 is good)**	**TLI (>0.9 is good)**	**RMSEA (<.08 is good)**	**SRMR (<.08 is good)**
Weighted Least Square Mean and Variance (WLSMV)					
Physical demand	Face-to-Face	0.915	0.830	0.327	0.162
	Telephone	0.954	0.908	0.227	0.127
	Web	0.947	0.893	0.409	0.131
Religiosity	Paper	0.999	0.998	0.142	0.020
Anxiety symptoms		0.993	0.987	0.091	0.045
Life satisfaction		0.999	0.996	0.120	0.025
Maximum Likelihood (ML)					
Religiosity	Paper	0.971	0.912	0.232	0.023
Anxiety symptoms		0.935	0.869	0.149	0.047
Life satisfaction		0.969	0.906	0.198	0.028

CFI: Comparative fit index, TLI: Tucker-Lewis Index, RMSEA: root mean square error of approximation, SRMR: standardized root mean square residual

12. Interviewer Variance in Key Measures

We present estimated interclass correlation coefficients (ICC)

- Based on simple multilevel (linear and logistic) models
- With no adjustments for other covariates (likely necessary, given the lack of interpenetration)
- Significant likelihood ratio tests indicate interviewer effects
- The magnitude of the interviewer-level clustering is relatively small

Key Variable	Estimated Iwer variance	Likelihood Ratio Test	ICC
Cognition score	0.50	136.4***	2.7%
CES-D symptoms	0.03	33.9***	0.9%
Self-rated Health	0.05	103.6***	1.5%
Nagi limitations	0.17	88.5***	1.4%
Work for pay?	0.19	84.7***	3.0%

Given that these are the “raw” ICCs, there would seem to be **very little evidence of interviewer variance** in these key measures (whatever the cause may be).

13. Straightlining and Speeding

Straightlining was assessed separately within two sections of the questionnaire: sections 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).

Straightlining is defined as:

- “strict” when the within-section SD = 0, indicating identical responses across all items
- “moderate” when within-section 0 < SD < 0.25, reflecting limited response variation

		Mode			
Section		**FTF**	**Telephone**	**Web**	**Total**
J (Employment, 14 Likert-type items)	Strict	0.1%	0.0%	0.4%	0.1%
	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%

Overall, the results present minimal evidence of straightlining behavior under both definitions.

13. Straightlining and Speeding

Speeding: Calculate an indicator of **unusually fast section completion.**

- Include IWs completed either face-to-face or by telephone with timing information (n = 11,655)
- Use section-level response times from the first **five** sections (asked of all Rs)
- Using method proposed by Leiner (2019)
  - For each section
    - Calculate a speed factor as the ratio **(median section completion time) : (R's completion time)**
    - Cap the speed factor at three
  - Averaged the capped speed factors across the five sections
  - A higher speed factor indicates faster-than-typical completion relative to other respondents.
- Classify respondents based on average speed factor:
  - slower than typical, slightly faster than typical, moderately faster than typical, and very fast

13. Straightlining and Speeding

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.

Classification	Criterion	Mode		
		Face-to-face (n=5595)	Telephone (n=5402)	Total
Slower than typical	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%

14. Breakoff Rate

In the HRS, interviews suspended prior to the delayed word recall test in section D are considered “breakoffs”. **The breakoff rate among panelists who started the 2022 HRS core survey is minimal: 0.6%.** This estimate includes all responses from all three modes.

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%)

15. Mode Effects

To evaluate potential mode effects, we **compared responses across face-to-face, telephone, and web interviews**.

The table below presents unweighted estimates of the 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.

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

Key takeaways:

1. Despite the significant results (given the large sample sizes), **effect sizes were extremely small.**
2. Web respondents had lower cognition and depression scores, reported higher self-rated health, and reported fewer Nagi limitations.

16. Survey Design

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

Survey content: 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.

Pre-tests: 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.

Proxy Rs: 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. Data Editing and Imputation

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.

Domain 3:
Sampling Error

18. Target Population
19. Achieved Sample Size

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.

The achieved sample size in the 2022 HRS panel was 13,128 interviews.

20. Design Effects and Coefficients of Variation

The design effects for the 2022 HRS panel 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.**

Variable	Weighted Estimate	Design-Adjusted Standard Error	Design Effect (DEFF)	Coefficient of Variation (CV)
Cognition	16.014	0.070	3.18	0.44%
CES-D symptoms	1.297	0.029	2.76	2.25%
Self-rated Health	0.413	0.008	3.59	2.06%
Nagi limitations	3.347	0.057	3.63	1.69%
Work for pay	0.367	0.008	3.53	2.25%

21. 1 + L / Unequal Weighting Effect

The overall unequal weighting effect (Kish's 1+L value) for the 2022 HRS panel is 2.03, indicating that variability in the 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. Sampling Description

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 HRS sample design can be found here.

23. Link to Analytic Guidelines

Analytic guidelines for users of the HRS data, including examples of syntax for performing analyses in statistical software packages with procedures available for the analysis of complex sample survey data, can be found here:

https://hrs.isr.umich.edu/documentation/new-user-guide

Overall Takeaway: From a holistic Total Survey Error perspective, the 2022 HRS panel is in extremely good shape, and data users can rest assured that they are working with high-quality survey data.

Additional Lessons Learned

Creating the first HRS DQP required a relatively large setup cost.

Lesson 4: 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.

Lesson 5: Subsequent DQPs created for other waves of HRS data collection will take advantage of the code that was created for this initial profile and therefore should take significantly less time to create.

Other projects are welcome to use the HRS code created for this process!

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.

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.

Questions?

We welcome any and all feedback.

Please direct questions to Brady (bwest@umich.edu) or Heather (schroedh@umich.edu).