

Machine-Learning Attrition Weights for the 2016–2020 General Social Survey Panel

David Crabtree

October 2026

Summary

In 2020 the General Social Survey (GSS), run by NORC at the University of Chicago, tried to reinterview people it had first interviewed in 2016 or 2018. Only 42% of eligible respondents took part. NORC’s released panel weight adjusts for that loss only by geographic area. In this research note, I evaluate various methods for generating attrition weights and compare them to the weights NORC provided.

I compare attrition weights from the standard logistic regression used in earlier GSS work to machine-learning methods: penalized regression, random forests, and gradient boosting. I find penalized logistic regression produces weights that remove the most bias. Because the people who left the panel still have their full baseline interview in the data, the right answer is known for any question asked before anyone left, so I can check whether each set of weights makes the people who stayed look like everyone who started. Panel studies have long used this kind of check (Fitzgerald, Gottschalk, and Moffitt 1998); to my knowledge it has not been applied to this panel’s weights. These findings stand to benefit GSS users whose questions involve answers that were affected by the panel’s attrition patterns.

Key Finding: Weights from a machine learning model that uses the whole questionnaire cut the attrition bias roughly in half; NORC’s released weight cuts it by about a tenth.

1. **Model-based weights beat both NORC’s weight and demographic raking.** Across 15 benchmark measures held out of every model, NORC’s weight closes 11% of the gap, raking on demographics 29%, and weights from the five attrition models 37%–53%.
2. **The panel kept more voters, more trusting people, and more liberals.** At baseline, the people who stayed were 7.0 points more likely than all eligible respondents to say most people can be trusted, 6.7 points more likely to have voted, and 3.9 points more likely to call themselves liberal. Dropout tracked education much more steeply than in the GSS’s 2006–2014 panels.
3. **The best predictor is not the best weight.** Random forests and gradient boosting ranked dropout best but removed less bias than penalized logistic regression. Choosing an attrition model by predictive accuracy alone would pick the wrong one.
4. **Better weights change 2020 estimates but not estimates of change.** Relative to NORC’s weight, they lower the 2020 share who say most people can be trusted by 3.6 points and the share with a great deal of confidence in science by 3.1 points. Within-person change from baseline to 2020 barely moves under any weighting.

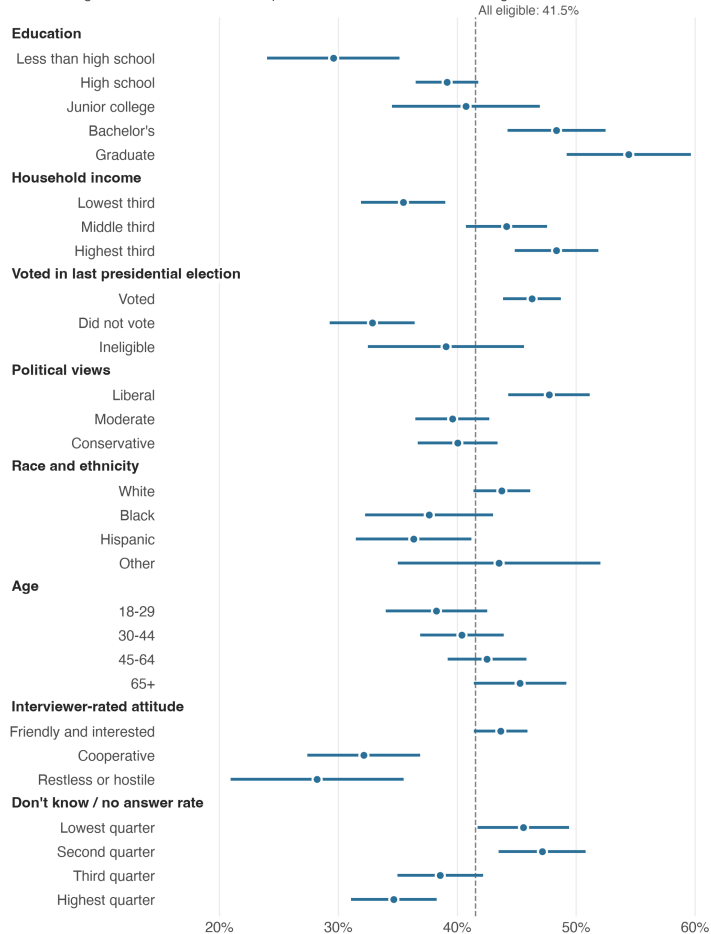
Who left the panel

The 2016–2020 GSS Panel reinterviewed every 2018 respondent and a random three-quarters of 2016 respondents, 4,494 people in all, from August to September 2020, mostly by web. After setting aside 143 who had died, moved abroad, or entered an institution, 4,351 were eligible and 1,823 completed the reinterview. Following Morgan and Lee’s (2020) work on the GSS’s earlier panels, I treat the eligible respondents as the population the weights should represent.

Who came back depended strongly on socioeconomic status and on how the baseline interview went. Among respondents without a high school diploma, 30% were reinterviewed; among those with a graduate degree, 54%. Respondents who had voted in the previous presidential election came back at 46%, nonvoters at 33%. Respondents whom the baseline interviewer rated “friendly and interested” came back at 44%, those rated restless or hostile at 28%.

Who came back in 2020

Share of eligible 2016 and 2018 GSS respondents reinterviewed in 2020, weighted, with 95% confidence intervals



The education gradient is much steeper than in the GSS’s 2006–2014 face-to-face panels: a graduate degree lowered the chance of dropping out by 15.3 points relative to a high school diploma, against 3.9 points then (Morgan and Lee 2020).

Predicting who left

I compared five models of who would be reinterviewed, each predicting from information collected at baseline:

- **Logistic regression with literature predictors:** about 20 predictors modeled on Morgan and Lee’s attrition models for the earlier panels (demographics, a vocabulary test, interviewer traits, and the interviewer’s ratings of the interview).
- **The same, plus principal components** of the other 355 questionnaire items, an unsupervised summary of respondents’ attitudes and circumstances.
- **Penalized (lasso or ridge) logistic regression, a random forest, and gradient boosting,** each using every baseline item asked in both years.

The 15 benchmark measures, and every item from the same question batteries, were held out of all models so they could be used to test the weights. Every respondent’s predicted probability comes from a model that never saw them (five-fold cross-validation, with tuning and preprocessing inside the training folds).

Model	AUC	Calibration slope
Logistic regression, literature predictors	0.597	0.72
Logistic regression + questionnaire components	0.604	0.70
Penalized logistic regression, all items	0.612	0.80
Random forest, all items	0.620	1.37
Gradient boosting, all items	0.619	1.11

AUC: how well a model ranks stayers above leavers (0.5 is chance). Calibration slope: 1 means predicted probabilities match observed rates; below 1, they are too extreme.

Dropout is only weakly predictable, as in the earlier panels (Frankel and Hillygus 2014; Morgan and Lee 2020). The tree models rank it best, improving AUC by 0.022 over the literature logistic regression (95% interval 0.007 to 0.037 for boosting), and boosting’s probabilities are the best calibrated.

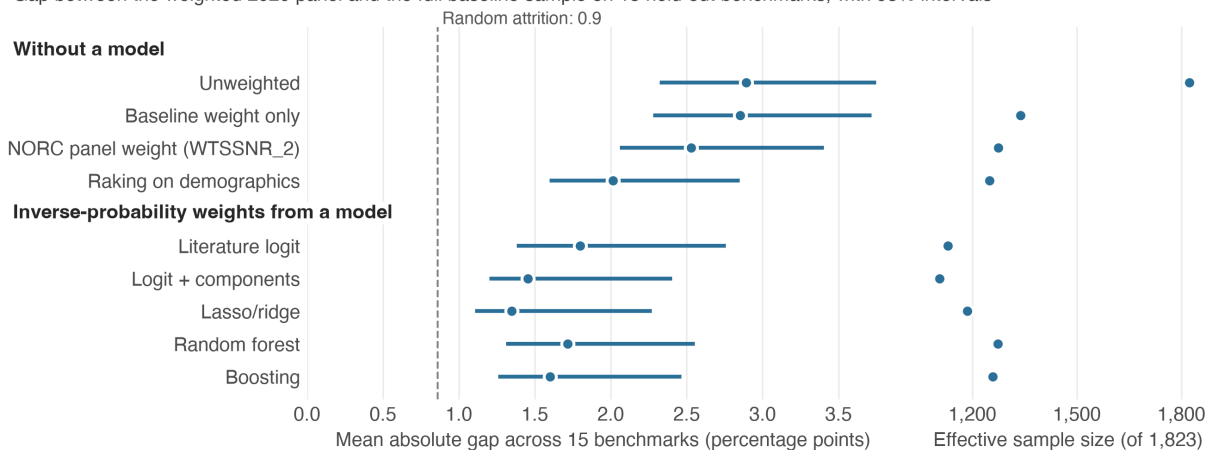
Do the weights remove bias?

Each model’s probabilities become weights by inversion: a respondent the model gave a 25% chance of returning stands in for four people like them, one with a 50% chance for two. I compared these with NORC’s released weight, with raking to the full sample’s age, sex, race and ethnicity, education, region, and baseline year, and with no attrition adjustment. (Trimming the largest weights or grouping respondents into propensity classes made little difference or did worse; both are in the output tables.)

For each held-out benchmark, the target is its baseline value among all eligible respondents, and each weighting scheme estimates it from the 1,823 who stayed. The difference is the attrition bias: what attrition adds to the panel’s estimate of a question asked before anyone left. (It is not error relative to the U.S. population, since the baseline sample has its own sampling error and nonresponse.) Even random attrition would leave an average gap of about 0.9 points, because those who stayed are a 42% subsample. Intervals come from 1,000 bootstrap resamples of the survey’s sampling units.

Model-based weights cut the attrition gap by 37-53%; NORC's panel weight by 11%

Gap between the weighted 2020 panel and the full baseline sample on 15 held-out benchmarks, with 95% intervals



Without attrition adjustment, the panel misses the baseline values by 2.85 points on average. NORC's weight lowers that to 2.53, a reduction its interval cannot distinguish from zero. Raking lowers it to 2.02. The penalized logistic regression on the full questionnaire does best, at 1.35: -1.18 points (95% interval -1.47 to -0.64) relative to NORC's weight and -0.67 points (95% interval -0.95 to -0.14) relative to raking. Its weights cost little precision: an effective sample size (Kish 1965) of 1,185, against 1,274 for NORC's weight.

Neither ranking nor calibration decided which weights worked. Gradient boosting predicted dropout best and was best calibrated, yet removed less bias (1.60) than the penalized regression; adding principal components barely improved the literature model's predictions but cut its remaining bias from 1.80 to 1.45. Weights reduce bias only to the extent that the attrition model captures characteristics related to the outcomes being estimated (Little and Vartivarian 2005). Here those are political and social attitudes, which the questionnaire-wide models see and the demographic and paradata predictors largely do not. Pew Research Center reached a similar conclusion when weighting online opt-in samples: adding political variables to demographic weighting reduced bias far more than any more complex statistical method did (Mercer, Lau, and Kennedy 2018).

What changes in 2020

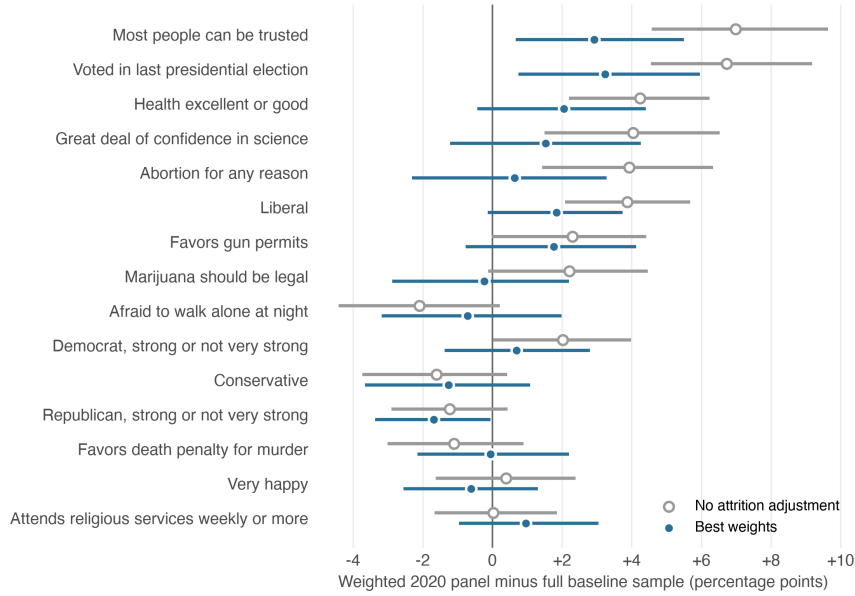
The benchmark questions were asked again in 2020. There is no true 2020 value to check against, but the bias test indicates which direction is right.

Relative to NORC's weight, the best weights lower the 2020 share who say most people can be trusted from 39.4% to 35.8%, the share with a great deal of confidence in science by 3.1 points (95% interval 1.8 to 4.5), and the share who voted in 2016 by 2.6 points (95% interval 1.6 to 3.5). Across the 15 measures, levels move by 1.3 points on average, while within-person change moves by only 0.4, with no shift distinguishable from zero: attrition bias largely cancels when the same people are compared with themselves. For users of this panel, the choice of attrition weight matters for 2020 levels and much less for change.

Limits. The bias test checks baseline answers, so weights that recover what panel members said in 2016 or 2018 may still miss differences that emerged by 2020. No weighting removes dropout tied to things never measured: trust and voting remain overstated by about 3 points even after

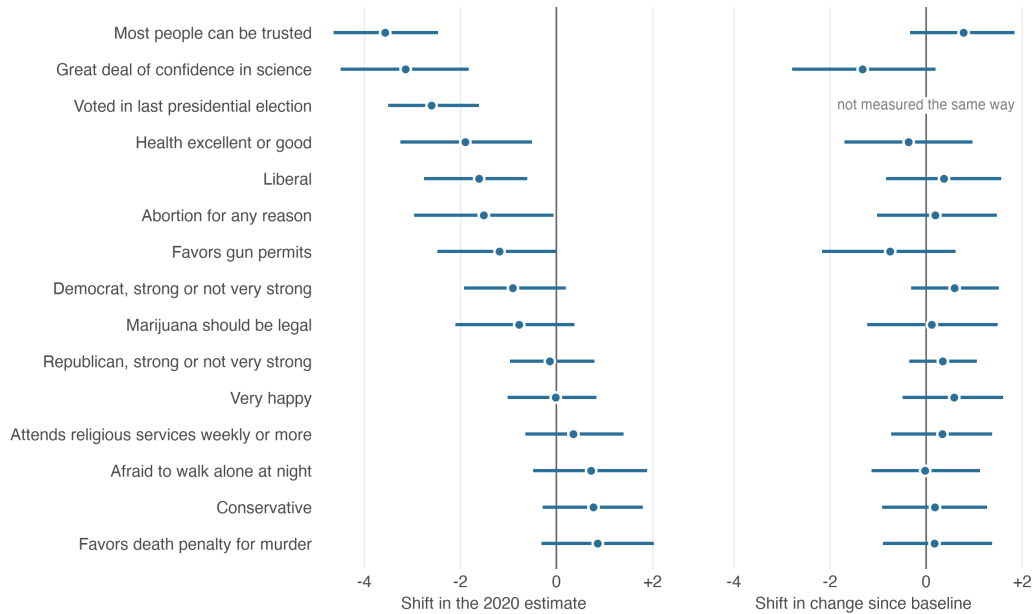
Who stayed: more voters, more trusting, more liberal

Baseline answers of 2020 panel members minus those of all eligible respondents, with 95% intervals.
Best weights: the scheme with the smallest mean gap (inverse probability, lasso/ridge).



The best weights move 2020 levels by up to 3.6 points but not change estimates

Inverse probability, lasso/ridge weights (best in the bias test) minus NORC's released weight, percentage points, with 95% intervals



the best weighting. The bootstrap holds each model’s predicted probabilities fixed, so its intervals understate uncertainty somewhat.

Data and code. The data are the public 2016–2020 GSS Panel (Release 1a), which the pipeline downloads from NORC; the repository’s `run_all.sh` reproduces every number and figure in this report.

References

Fitzgerald, J., Gottschalk, P., and Moffitt, R. (1998). An analysis of sample attrition in panel data: The Michigan Panel Study of Income Dynamics. *Journal of Human Resources*, 33(2), 251–299.

Frankel, L. L., and Hillygus, D. S. (2014). Looking beyond demographics: Panel attrition in the ANES and GSS. *Political Analysis*, 22(3), 336–353.

Kish, L. (1965). *Survey Sampling*. Wiley.

Little, R. J. A., and Vartivarian, S. (2005). Does weighting for nonresponse increase the variance of survey means? *Survey Methodology*, 31(2), 161–168.

Mercer, A., Lau, A., and Kennedy, C. (2018). *For Weighting Online Opt-In Samples, What Matters Most?* Pew Research Center. <https://www.pewresearch.org/methods/2018/01/26/for-weighting-online-opt-in-samples-what-matters-most/>

Morgan, S. L., and Lee, J. (2020). *Six Alternative Weights that Adjust for Attrition in the 2006–2014 General Social Survey Panels*. GSS Methodological Report No. 132. NORC at the University of Chicago.

Data. Davern, M., Bautista, R., Freese, J., Morgan, S. L., and Smith, T. W. *General Social Survey 2016–2020 Panel*, Release 1a. NORC at the University of Chicago. <https://gss.norc.org>