ASSIGNMENT ERROR BY BALLOT AND FORM

FINAL METHODOLOGICAL REPORT #134, OCTOBER 2022

Authors: Leah Christian, Nicholas Davis, Caroline Lancaster, and Susan Paddock[1]

CONTENTS

Executive Summary 3
Overview of the Assignment Error 4
Weights to Address the Assignment Error 7
Analysis of Differences Between Original and Adjusted Estimates 10
Discussion and Conclusion 22
Methodology 23
References 26
Appendix 27

[1] The authors wish to acknowledge our key collaborators: Erika Archie, Nola du Toit, Amy Ihde, Ji Eun Park, Joshua Lerner, Kiegan Rice, Erik Scherpf, and Andrew Stern. We would also like to thank our reviewers: René Bautista, Lisa Blumerman, Mike Davern, Jeremy Freese, Pamela Herd, Steve Morgan, and Eric Young. Judith A. Seltzer and Megan M. Sweeney uncovered this issue in the course of their research. Michael Hout reported it to the GSS PIs and Board after Judith Seltzer showed him their findings. Judith Seltzer and Nora Cate Schaeffer provided valuable feedback on the preliminary analysis and report.

List of Exhibits

Exhibit 1: Theoretical Illustration of Kish Roster Number Selection and Assignment Error by Ballot and Form 6

Exhibit 2: Household Member Generally Assigned to Ballot in 3-Adult Households and Form in 2-Adult Households in the Affected Years........ 7

Exhibit 3: Original and Adjusted Proportion 18–49 by Form (2000–2018) 9

Exhibit 4: Original and Adjusted Proportion Female by Form (2000–2018)........ 10

Exhibit 5: Woman Will Not Get Promotion (2018) 11

Exhibit 6: Avg. Percentage Point Difference between Original and Adjusted Estimates across Categorical Questions 12

Exhibit 7: Avg. Difference across Categories between Original and Adjusted Estimates (2018) 12

Exhibit 8: Avg. Difference on Ballot and Form Subsets........ 13

Exhibit 9: Avg. Percentage Point Difference, Current Affairs (2018) 14

Exhibit 10: Avg. Percentage Point Difference, Civil Liberties (2018)........ 14

Exhibit 11: Number of Categories and Questions with a Potentially Meaningful Difference 15

Exhibit 12: Variables with a Potentially Meaningful Difference in Estimates (2018) 16

Exhibit 13: NATENVIY, National Spending on the Environment (2018) 17

Exhibit 14: NATEDUC, National Spending on Education (2018) 17

Exhibit 15: POSSLQ, Marital and Cohabitation Status (2018)........ 17

Exhibit 16: BIGBANG, The Universe Began with a Huge Explosion (2018) 18

Exhibit 17: GENDER2, Gender of the Second Person in the Household (2018)........ 18

Exhibit 18: Ordered Logistic Regression, Original and Adjusted Weights (NATENVIY, 2018).... 19

Exhibit 19: Logistic Regression, Original and Adjusted Weights (CAPPUN, 2018) 20

Exhibit 20: Ordered Logistic Regression, Original and Adjusted Weights (FECHLD, 2018)........ 21

Exhibit 21: Logistic Regression, Original and Adjusted Weights (OWNGUN, 2018)........ 22

Executive Summary: The General Social Survey (GSS) has three ballots and two forms that combine to create six different versions of the survey questionnaire. This approach enables us to reduce respondent burden by asking some questions of subsamples of respondents and also allows us to conduct question wording experiments using the two forms. The GSS also randomly selects an adult in the household to complete the survey interview to ensure that the probability sample design is extended to selecting a respondent within the household. An error occurred in how respondents were assigned to the ballots and forms of the GSS questionnaire in 2002, 2010, 2012, 2016, and 2018, resulting in imbalances across questionnaire versions in some types of households. This error does not affect the random selection of the respondent within household in these years. However, the error resulted in the mean age varying by questionnaire ballot and form, as well as associated demographic characteristics that are patterned by the age structure of households (e.g., sex). GSS users helped to identify the error by notifying NORC they had discovered demographic variations by form. Importantly, for questions asked on all ballots and forms of the GSS, the assignment error did not impact the composition of the sample or the estimates.

NORC created new weights that address the assignment error, and this report summarizes our analysis[2] of the assignment error. The analysis compares the estimates for all questions in the affected years using the original and the new weights to approximate the bias from the error. The average difference across categories between the original and adjusted estimates is 0.42 for questions that are asked on only some forms or ballots in the affected years (and 0.36 percentage points across all questions). Fully 95.0 percent of questions tested that were asked on some forms or ballots had an average difference across categories of 1 percentage point or fewer and only 0.2 percent (4 questions) had a difference of 3 percentage points or greater. Furthermore, equivalence testing indicates that potentially meaningful differences exist between the two sets of estimates for only 1.2 percent of questions (24 questions) across the affected years.

This report reviews the findings from the analysis, and the appendix includes a detailed breakdown of all questions in the affected years, the magnitude of the difference, and whether the difference was statistically meaningfully different. The consequences of the assignment error by ballot are small because the error was present only for 3-adult and 6-or-more adult households (10.9 percent of respondents in 2018), and most questions appear on at least two of the three ballots. As a result, the error was partially or fully offset in the aggregate because of the variation in age structure that exists across the different households impacted, and misassignment in the affected households is minimized when data across ballots are aggregated.

For the assignment error by form, more households were impacted because the error was present for 2-adult, 4-adult, and 6-or-more adult households (55.5 percent of respondents in 2018). Since more households were impacted, this resulted in larger age imbalances by form in

[2] This methodology report is updated with our final analysis of the assignment error by ballot and form; the preliminary analysis is archived here.

the years affected by the assignment error. The question wording experiments that use form had the most potential to be impacted by the assignment error, given the number of households impacted. Nonetheless, the analysis shows that the impacts are small for most questions asked on only one form, although a few questions do show larger impacts.

The overall conclusion of the report is that the impact of the error was partially or entirely self-correcting and thus has minimal consequences for analysis of GSS data, including those published in the past. Overall, our results indicate that nearly all substantive conclusions based on simple associations will be the same whether analysts use the original or the new weights. While we cannot rule out the possibility that users' multivariate analyses may be affected, our example analysis suggests that any differences should be minimal. Even so, the new weight should be used in most cases when analyzing questions that are asked on only some ballots or forms in the affected years.

Overview of the Assignment Error

Selection of Eligible Adult within the Household: The GSS selects one randomly selected adult in the household to complete the survey interview. To ensure that all eligible adults in the household have a known, nonzero chance of being selected as the respondent, the GSS uses a sampling procedure known as a Kish methodology (Kish, 1949). This method consists of creating a roster of all adults in the household listed in order from oldest to youngest and then selecting an eligible adult from the roster using Kish numbers based on a predetermined sequence that has a random start. In a household with only 1 eligible adult, the first household member is always selected (the only eligible adult is always selected from the household roster). In 2-adult households, the Kish numbers are the first or second adult listed on the household roster (e.g., each adult has a 50 percent chance of being selected). In 3-adult households, the Kish numbers are the first, second, or third adult listed on the household roster, and the process continues similarly for larger households.[3] The Kish numbers are assigned to each household when the sample is drawn using systematic sampling, and the random starting point for the Kish digits are randomized differently for each GSS year (for instance, the random start for Kish numbers in 2018 is different from 2016 or other previous years). This cycle repeats, assigning Kish numbers of 1 to N, where N is the number of eligible adults in the household.

Assignment of Questionnaire to the Selected Respondent: There are three ballots (ballots A, B, and C) and two forms (forms X and Y) that combine to create six different versions of the GSS questionnaire in 2018 (the other affected years have a similar configuration). These three ballots are used to help reduce respondent burden so that not all questions are asked of the full sample. Most of the core GSS attitude and opinion questions are included on all three ballots or two of the three ballots. There are also two forms, which are used to conduct experimental research for a small subset of attitudes and opinions. These X and Y forms are usually

[3] In the GSS, any households with 6 or more eligible adults are capped at 6 for this respondent selection process. The Kish process selects from the six oldest adults in the household.

introduced to enable question-wording experiments, but they have been used for additional purposes as well.

A systematic random procedure is used to assign questionnaire ballot and form to the sampled respondent. For this process, a version of the questionnaire (combination of ballot and form) is preassigned to every sixth sample household (i.e., based on a regular interval with a random start). A sequence of 6 is used for the simultaneous allocation of form and ballot, and this sequence has a cycle 3 for ballot (A, B and C) and cycle 2 for form (X and Y).[4]

DESCRIPTION OF THE ASSIGNMENT ERROR: In late 2021, GSS users identified in historical data a sex difference by form that appeared to affect multiple years of the GSS. The NORC team investigated this difference and verified that an error occurred in how respondents were assigned to form and ballot. The error in the affected years is that the assignment of Kish numbers, questionnaire form, and questionnaire ballot had the same sort order for all three assignments. This error caused an unintended relationship between the age order of adults within household and the questionnaire ballot and form assigned for households of certain sizes in 2002,[5] 2010, 2012, 2016, and 2018, as detailed below. In all years, the selection of the eligible adult was correctly implemented.

The assignment error resulted from the fact that the age order of adults in the household determined the assignment of ballot and form to the selected respondent (in households with multiple eligible respondents). Exhibit 2 provides a simplified example of the Kish methodology and questionnaire assignment to illustrate the error. Since there are three ballots, the error affected 3-adult and 6-or-more-adult households. In 3-adult households in this example, the oldest adult (#1) always would be assigned to ballot A, the middle adult (#2) to ballot B, and the youngest adult (#3) to ballot C. Thus, respondents who were assigned ballot A were systematically older than those who were assigned ballot C. As discussed further in the Impact on Demographics section, although the assignment error was dependent on age order, it impacted other demographics related to age order and household composition in addition to impacting the age distributions. This pattern is repeated in 6-or-more-adult households, but where the oldest and the fourth oldest always get ballot A, and so on. The assignment error by ballot in 2018 affected the 10.9 percent of respondents living in households of these sizes.

[4] To ensure that the questionnaire allocation is balanced across geographies, an implicit stratification is used to reduce sampling variance and the selected households are geographically sorted (on sample segments) with the purpose of allocation of ballot and form; consequentially, form-ballot combinations are distributed uniformly within geographic segments.

[5] In 2002, the assignment error only impacted the assignment of ballot but not form, and thus only 3-adult and 6-or-more adult households were impacted by the error.

Exhibit 1: Theoretical Illustration of Kish Roster Number Selection and Assignment Error by Ballot and Form

	KISH ROSTER SELECTION						QUESTIONNAIRE ASSIGNMENT	
SAMPLE HOUSEHOLD	ONE ADULT	TWO ADULTS	THREE ADULTS	FOUR ADULTS	FIVE ADULTS	SIX+ ADULTS	BALLOT	FORM
HOUSEHOLD 1	1	1	1	1	1	1	A	X
HOUSEHOLD 2	1	2	2	2	2	2	B	Y
HOUSEHOLD 3	1	1	3	3	3	3	C	X
HOUSEHOLD 4	1	2	1	4	4	4	A	Y
HOUSEHOLD 5	1	1	2	1	5	5	B	X
HOUSEHOLD 6	1	2	3	2	1	6	C	Y
HOUSEHOLD 7	1	1	1	3	2	1	A	X
HOUSEHOLD 8	1	2	2	4	3	2	B	Y
(…)	(…)	(…)	(…)	(…)	(…)	(…)	(…)	(…)

The nature of the assignment error by form was similar but since there are two forms, the error affected 2-adult, 4-adult, and 6-or-more-adult households. For example, in 2-adult households, the oldest adult (#1) would always be assigned to form X and the youngest adult (#2) would always be assigned form Y. Thus, respondents who were assigned form X were systematically older than those who were assigned form Y. This pattern is repeated in 4-adult households, but where the oldest (#1) and third oldest (#3) would always be assigned form X, and the second oldest (#2) and the youngest (#4) would always be assigned form Y. The assignment error by form impacted a larger number of households; in 2018 it affected the 55.5 percent of respondents living in 2-adult, 4-adult, and 6-or-more-adult households.

Because the starting point for the Kish numbers is randomized differently for each GSS year, the ballot and form assignment to household members differed in the affected years. To help illustrate how the error manifests differently in the affected years, Exhibit 2 shows the general pattern of which household member (oldest, middle, youngest) was assigned to a ballot in 3-adult households and which household member (oldest, youngest) was assigned to a form in 2-adult households for each affected year. See Appendix A for a more detailed table of the assignment error by ballot and form in 2018.

Exhibit 2: Household Member Generally Assigned to Ballot in 3-Adult Households and Form in 2-Adult Households in the Affected Years

	3-Adult Households			2-Adult Households	
	Ballot A	Ballot B	Ballot C	Form X	Form Y
2018	Middle adult	Oldest adult	Youngest adult	Youngest adult	Oldest adult
2016	Oldest adult	Youngest adult	Middle adult	Youngest adult	Oldest adult
2012	Middle adult	Youngest adult	Oldest adult	Youngest adult	Oldest adult
2010	Middle adult	Youngest adult	Oldest adult	Oldest adult	Youngest adult
2002	Youngest adult	Oldest adult	Middle adult	No error	

Weights to Address the Assignment Error

Description: Two new weights were created to adjust for the assignment error by ballot and form and allow for examining the impact of the error on GSS estimates. The new weights are the rebalanced analogs of the existing GSS weights, WTSSALL and WTSSNR.[6] Each version of the weight was rebalanced in the affected years to adjust for the assignment error by ballot and form in the households impacted and set equal to the original version of the weight in unaffected years since these years had no assignment error and therefore no correction was needed.

To accomplish this rebalancing, respondents in households affected by the assignment error were isolated, and weights were only updated for households affected by the error (2-adults, 3-adults, 4-adults, and 6-or-more adults). Then, the existing weights among respondents in these households were adjusted using a raking methodology. We performed the raking within combinations of ballot and form, and include the dimensions of sex, age, marital status, and work status — variables determined to be skewed due to the assignment error. The resulting weighted demographic estimates within each combination of ballot and form are equal to the weighted estimates for the full set of respondents in affected households.

Separately from these analyses, the GSS team developed post-stratified weights for GSS Cross-section for 2000 through 2018 (these are in addition to previously released post-stratified weights for 2021 GSS Cross-section). Those weights include the rebalancing adjustments described here in addition to additional adjustments for demographic characteristics to account

[6] These new weights (BALLOTFORMWT and BALLOTFORMWTNR) will be released with Release 3 of the 2021 GSS cumulative data file. In addition, Release 3 includes the original unadjusted weights WTSS, WTSSALL, and WTSSNR, so interested researchers can conduct their own investigations into the impact of the error on their specific analyses.

for differential nonresponse based on U.S. Census Bureau data.[7] We recommend users employ those weights in practice moving forward to adjust for the assignment error and nonresponse.

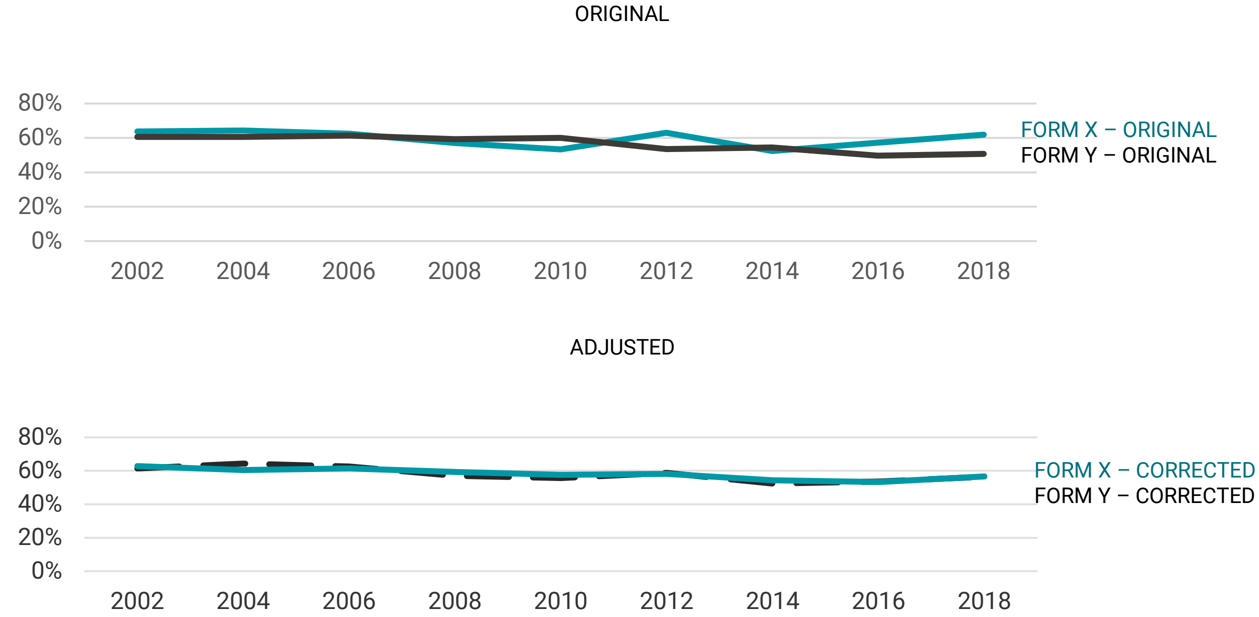
Throughout this report, we used the adjusted version of the WTSSALL weight since the analytic results in this report are focused on the impact of the assignment error and how the new weights adjust for it, separate from nonresponse adjustments. In general, the new weights produce modest increases in estimated design effects and simulated margins of error for hypothetical estimates, meaning that using the adjusted weights will result in minimal loss of precision in estimation (see Appendix B for more detail on the coefficient of variation, design effects, and simulated margins of error for the original and new weights).

IMPACT ON DEMOGRAPHICS: To examined how the adjusted weights impacted the demographic characteristics of the GSS sample, we compared the age and sex estimates of GSS respondents derived from the original and adjusted weights for the affected years. The analyses indicate that adjusting for the assignment error resulted in very small changes to the age and sex estimates for the full sample in the affected years. This was expected since the imbalances only impacted individual ballots and forms (detailed tables showing age and sex for the total sample and by ballot and form are shown in Appendix C).

Since the weighting adjustments were performed on ballot and form combinations, we examined age by form and ballot in the affected years. As expected, the proportion 18 to 49 improved with the new weights, bringing the proportion on form X and Y closer in the affected years (as shown in Exhibit 3). Similarly for ballot, the proportion 18 to 49 and 50 and older improved with the new weights, with the proportions closer to the full sample and more balanced by ballot (as shown in Appendix C).

[7] See here for more details about these new post-stratified weights for GSS 2000–2018 Cross-section and plans for creating post-stratified weights in earlier years of the GSS.

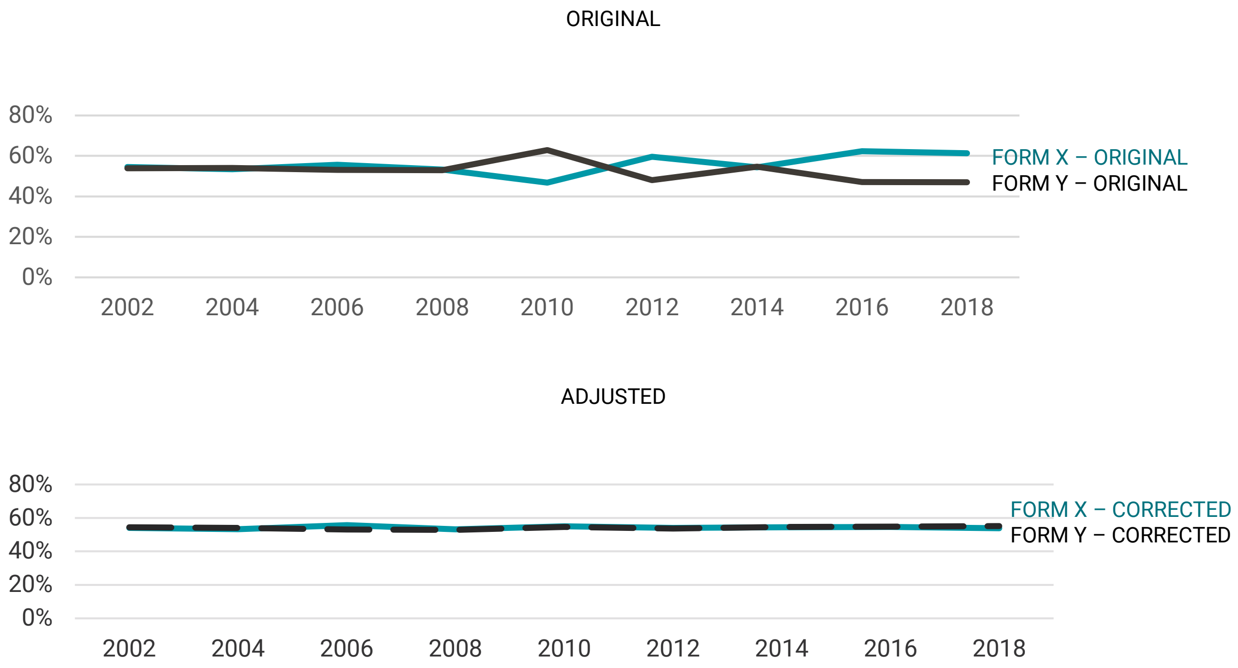
Exhibit 3: Original and Adjusted Proportion 18–49 by Form (2000–2018)



The errors in assigning form and ballot can also affect the sex composition of the sample responding to each ballot or each form. For example, in 2-adult households, the older of the two adults is disproportionately male (61.7 percent of the oldest adults in 2018 were male) whereas the youngest of the two adults is more likely to be female (68.8 percent of the second-youngest adults in 2018 were female). Although less severe, the assignment error can affect other demographic characteristics associated with the age rank, such as marital status and employment status. For example, the oldest person in the household is most likely to be employed. Similarly, in 4-adults and 6-or-more-adult households, the oldest person in the household is likely to be married. Acknowledging the complex relationship between age rank and other demographic characteristics, we adjusted for age, sex, marital status, and work status in the reweighting.

To understand how the adjusted weights affected the balance in sex for the affected years, we examined the proportion of female and male by form and ballot. The adjusted weights created a more balanced distribution of females and males for each form (as shown in Exhibit 4). For example, in 2018, 61.3 percent of respondents assigned to form X were female, while 38.7 percent were male. The adjusted weights improved the balance by changing the balance to 53.8 percent female and 46.2 percent male assigned to form X. The imbalance in sex by ballot was less severe (55.1 percent female on ballot A, 54.2 percent on ballot B, and 53.4 percent on ballot C), and the adjusted weights only had a small impact on the proportion of female and male by ballot (shown in Appendix C). Differences between the original and adjusted estimates were less than 2 percentages points but generally move the estimates closer to the overall distribution of GSS respondents in 2018.

Exhibit 4: Original and Adjusted Proportion Female by Form (2000–2018)



Analysis of Differences Between Original and Adjusted Estimates

Description of Approach: We conducted analyses on all GSS questions[8] in the affected years to compare estimates derived from the original and new weights. We also present results for the questions asked on only some forms or ballots. We first conducted a difference analysis comparing estimates with the original and new weights. For categorical variables, we calculated the difference between the original and adjusted estimate for each category and then calculated the average absolute value of the differences across categories for each question.

Exhibit 5 presents an example of our categorical difference calculations, using DISCAFFW — the likelihood that a qualified woman will be passed over for a promotion — which was fielded on form Y of ballots A and B. The first two columns show the estimates for each category using the original and new weights, respectively. We then subtracted the adjusted estimate from the original estimate and took the absolute value of the difference (the absolute differences for this question ranged from 0.2 to 0.5 percentage points). The variable-level difference estimate is created by averaging over the category-level absolute differences (the average difference for this question was 0.4 percentage points). We report data for all affected years throughout the

[8] We excluded questions with fewer than 25 responses.

report, but we use data from 2018 to help explain and illustrate the impact of the assignment error for specific areas and questions.[9]

Exhibit 5: Woman Will Not Get Promotion (2018)

	ORIGINAL	ADJUSTED	ABS DIFF
VERY LIKELY	26.4%	26.2%	(0.2%)
SOMEWHAT LIKELY	45.2%	45.4%	(0.2%)
SOMEWHAT UNLIKELY	19.1%	19.7%	(0.5%)
VERY UNLIKELY	9.2%	8.7%	(0.5%)
AVERAGE DIFF			**0.4%**

For continuous variables, we calculated the difference between the mean of the original and adjusted estimate for each question. We also performed statistical equivalence testing on all questions to determine whether we statistically reject the presence of meaningfully large differences between the original and adjusted estimates. Additional details about the analyses can be found in the Methodological section.

The results of these analyses indicate that the impact of the assignment error is minimal. For example, 95.0 percent of questions asked on some forms or ballots that were tested in the difference analysis have an average difference of 1 percentage point or fewer across categories, and only 0.3 percent have an average difference of 3 percentage points or greater across the impacted years (the results are nearly identical for all questions tested: 94.9 percent and 0.3 percent, respectively). Further, equivalence testing indicates that meaningfully large differences between the original and adjusted estimates could not be ruled out for only 1.2 percent of questions. Our examination of these potentially meaningful differences suggests that substantive conclusions generally remain unaffected when using the new weights. In the sections below, we discuss each analysis in more detail and present key findings. In Appendix D, we share the full results from our analysis, including the average difference in proportions (discrete) or difference in means (continuous) for all variables and whether the question had potentially meaningful differences based on the statistical equivalence testing.

OVERVIEW OF DIFFERENCES: The difference analysis examined the absolute difference in response category proportions between estimates under the original weights and the adjusted weights. The average difference across categories between the original and corrected estimates is small — 0.36 percentage points across all questions in the affected years. As Exhibit 6 indicates, the average percentage point difference ranges from 0.24 percentage points in 2002 to 0.45 percentage points in 2018. The same is true for the median difference, which

[9] We chose 2018 since it was the most recent year impacted and because the analysis indicated it had somewhat larger differences than other impacted years.

ranges from 0.17 percentage points in 2002 to 0.33 percentage points in 2018. If we exclude questions asked on all ballots and forms, the average difference in each year is still less than half a percentage point, with an average difference across the affected years of 0.42 percentage points.

Exhibit 7 presents the distribution of differences for 2018, where 92.3 percent of questions have an estimated difference of 1 percentage point or fewer. Only five questions (0.6 percent) have an estimated average difference across categories of 3 percentage points or greater. These variables have a generally low number of observations each (ranging from 41 to 311[10]) and are not commonly used items. These results for 2018 are very similar for questions asked only on some ballots or forms, with 91.8 percent of questions having a difference of 1 percentage point or fewer, and three questions having a difference of 3 percentage points or greater. In other years, even fewer variables have a greater than 3 percentage point difference: zero variables in 2012; one in 2002 and 2010; and three in 2016.

Exhibit 6: Avg. Percentage Point Difference between Original and Adjusted Estimates across Categorical Questions

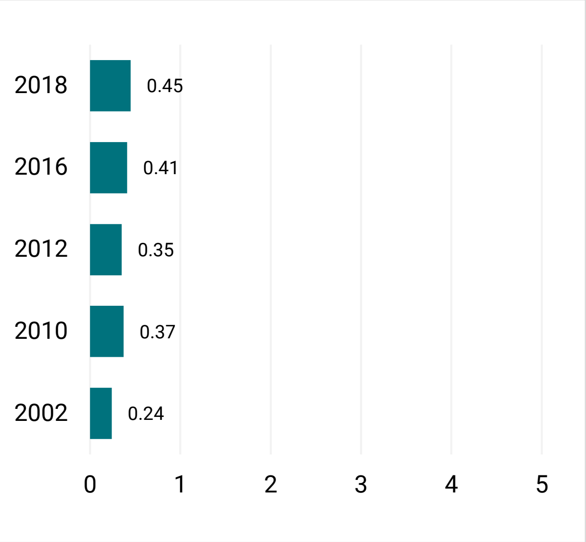
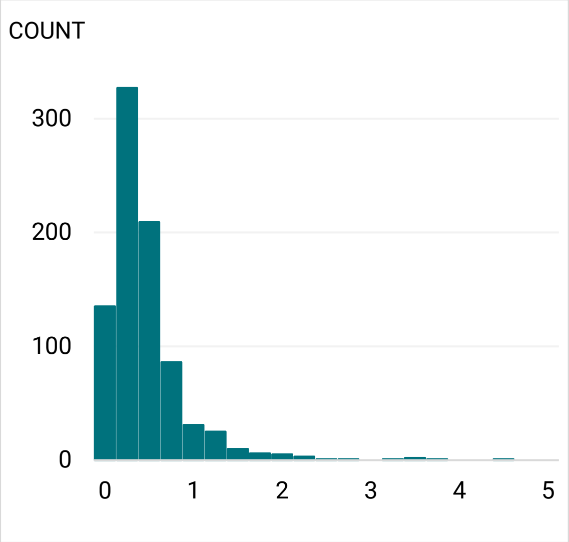


Exhibit 7: Avg. Difference across Categories between Original and Adjusted Estimates (2018)



We also examined the average difference across categories between the original and adjusted estimates for questions asked only on one ballot (A, B, or C); two ballots; or one form (X or Y), and the results are very similar. As shown in Exhibit 8, the average difference across these subsets is 0.42. As expected, the largest differences are seen with questions asked on only one

[10] FRNDSEX (N = 197); GENDER6 (N = 53); JEW16 (N = 41); KNWMW5 (N = 94); and WEBMOB (N = 311). Of these, FRNDSEX, KNWMW5 and WEBMOB were asked on only some ballots or forms.

ballot (average difference 0.65) or one form (average difference 0.54), and differences are attenuated for questions asked on two ballots.

Exhibit 8: Avg. Difference on Ballot and Form Subsets

	QUESTIONS ON ONE BALLOT	QUESTIONS ON TWO BALLOTS	QUESTIONS ON ONE FORM	TOTAL[11]
2018	0.62	0.47	0.74	0.49
2016	0.74	0.37	0.66	0.42
2012	0.54	0.42	0.54	0.39
2010	0.46	0.42	0.47	0.42
2002	n/a	0.31	0.19	0.29
TOTAL	0.65	0.41	0.54	0.42

Further, in Exhibits 9 and 10, we present the average difference across categories for 2018 questions in the “Current Affairs” and “Civil Liberties” GSS topic areas. These differences are generally small — only a handful of these variables have an average difference greater than 1 percentage point. Similar figures for the “Religion and Spirituality,” “Politics,” “Quality of Working Life,” and “Gender and Marriage” topic areas can be found in Appendix D.

[11] This column contains the average difference for questions not asked on all six questionnaires. The “total” column does not necessarily equal the average of the other three columns since questions may fall into more than one category (e.g., A question asked only on AX falls into both “one ballot” and “one form” categories).

Exhibit 9: Avg. Percentage Point Difference, Current Affairs (2018)

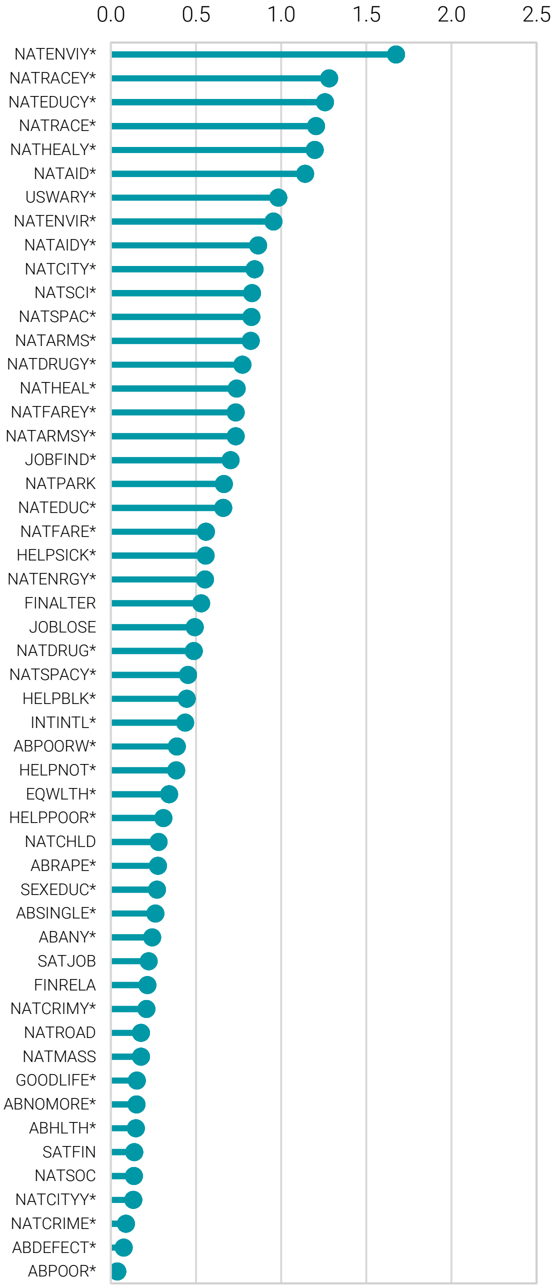
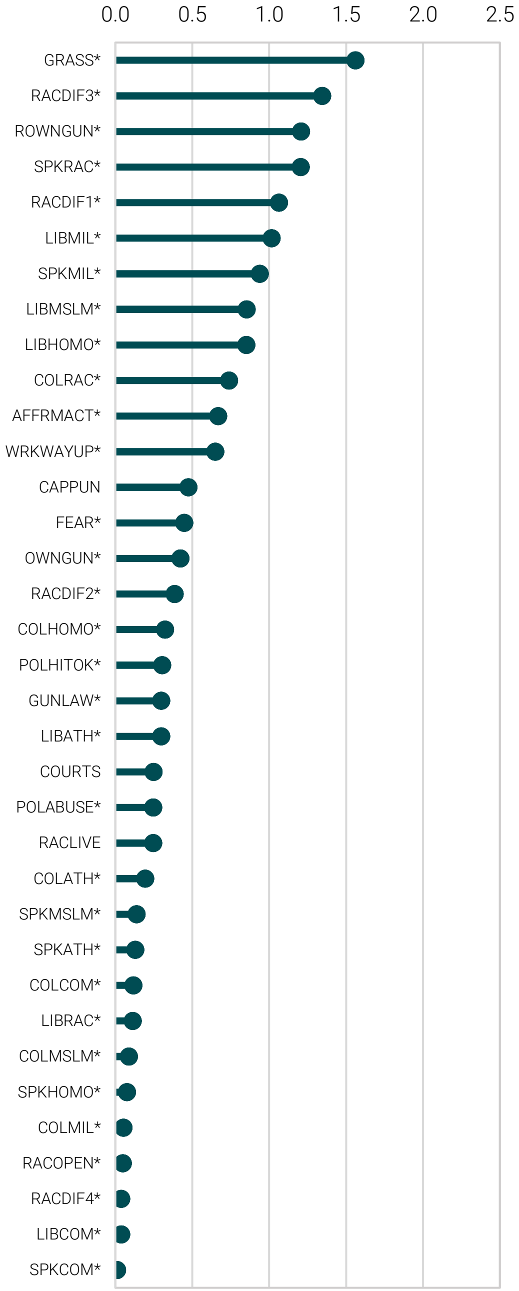


Exhibit 10: Avg. Percentage Point Difference, Civil Liberties (2018)



Note: Asterisks (*) in the above figures indicate questions that were asked only on some forms or ballots.

Statistical Testing: To test for substantive impacts to estimates, we employ equivalence testing of estimates with the original versus adjusted weights. The equivalence test requires hypothesizing that there is a meaningfully large difference between the original and corrected estimate and rejecting that hypothesis if the data show otherwise. This is the reverse of the more commonly used statistical testing approach that requires hypothesizing a zero difference and rejecting that hypothesis if the data show otherwise. We cannot use the common testing approach to conclude that the original and adjusted estimates are statistically similar — only that they are significantly different. We therefore do not conduct testing under null hypotheses of zero differences since we would expect the assignment error to impact these estimates.

The null hypothesis under an equivalence test asserts that the difference is greater than or equal to a predetermined "smallest effect size of interest" (SESOI) (Lakens et al., 2018). Two one-sided hypothesis tests (TOST) are performed to determine whether the difference lies outside upper and lower equivalence bounds, -Δ and Δ, defined by the SESOI. If both tests are rejected at the chosen α level, (in our case, 5 percent), then the estimates are declared statistically equivalent at that α level. We define the equivalence bounds for our analysis as the margin of error of the category mean (for discrete variables) or variable mean (for continuous variables) under the original weight. Testing was undertaken at the category-level for discrete variables and at the variable-level for continuous variables. Categories were included for testing if they contained more than 25 observations.

Exhibit 11: Number of Categories and Questions with a Potentially Meaningful Difference

	CATEGORIES FOR ALL QUESTIONS	ALL QUESTIONS	CATEGORIES ON ONLY SOME BALLOTS/FORMS	QUESTIONS ON ONLY SOME BALLOTS/FORMS
2018	**46** of 2,938	**42** of 965	**25** of 1,760	**21** of 545
2016	**6** of 2,765	**6** of 857	**3** of 1,771	**3** of 553
2012	**0** of 2,261	**0** of 743	**0** of 1,312	**0** of 392
2010	**3** of 2,506	**3** of 763	**0** of 1,253	**0** of 363
2002	**0** of 3,497	**0** of 1,007	**0** of 463	**0** of 160

As shown in Exhibit 11 above, the null hypothesis of a potentially meaningful difference in estimates was rejected for most questions, meaning that generally the original and adjusted estimates were statistically equivalent. Focusing only on questions asked on some ballots or forms, only 24 questions showed potentially meaningful differences. In fact, all questions in 2002, 2010, and 2012 have statistically equivalent estimates under the original and new weights. Even in 2018, the year with the highest number of potentially affected variables, only 21 questions (3.9 percent) of tests failed to establish statistical equivalence. Of these variables, 8 of the 21 are attitude/opinion questions, 5 are behavioral questions, and 8 are household, demographic, or technical variables. A list of these variables can be found in Exhibit 12.

Variables asked on all ballots and forms should theoretically not be affected by the assignment error, but we present the results for all questions as well so users can evaluate the impact of the adjusted weights. The fact that the adjusted weight produces estimates that are not statistically equivalent may be due to factors unrelated to the error. For example, the adjusted weights could introduce bias to variables that are correlated with gender but were not affected by the assignment error.

Exhibit 12: Variables with a Potentially Meaningful Difference in Estimates (2018)

QUESTION TOPIC	VARIABLE NAME (ALL)	VARIABLE NAME (ONLY ON SOME BALLOTS/FORMS)
ATTITUDES AND OPINIONS	none	CHARACTR, GOVLAZY, INTLWHTS, MCSDS4, NATEDUC, NATENVIY, PILINGUP, WORKHSPS
BEHAVIORS	LEARNNEW, SUPHELP, WKFREEDM	CONWKDAY, DWELOWN, HLPADVCE, PARTNERS, PARTPART
DEMOGRAPHICS, HOUSEHOLD, AND TECHNICAL	ETH2, GENDER2, GRANBORN, HHTYPE, HHTYPE1, INDUS10, INTHISP, ISCO08, ISCO88, MAOCC10, NUMEMPS, OCC10, OWNSTOCK, RELATE3, RELHH2, RELHH3, RELHHD2, RELHHD3	HLTHSTRT, KNWCLENR, KNWHRMAN, NATVIEWS, OTHMHNEG, RHLTHEND, SEXORNT, WORDB

Example Differences: In the following section, we provide several examples from 2018. The selected variables (in Exhibits 13–17) represent a variety of ballot and form combinations and illustrate a range of average difference sizes. Two variables, NATENVIY and NATEDUC, were flagged in our equivalence testing as having potentially meaningful differences, which is not surprising given these national spending questions are asked on only one form of each ballot and are part of the question wording experiments that use form X and Y versions of the questionnaire.

For NATENVIY, we see a relatively sizable average absolute difference of 1.7 percentage points. Yet only the difference of 2.5 percentage points for the “too little” category were flagged in our equivalence testing. These differences nonetheless do not affect the overall conclusions that most respondents believe the country is spending “too little” on the environment. (See Exhibit 13.)

NATEDUC has a smaller average absolute difference of 0.7 percentage points, with the 0.9 percentage point difference in the “too much” category deemed potentially meaningful. The prevalence of this response was slightly underestimated with the original weight, but it remains the least common response under the new weight. (See Exhibit 14.)

Exhibit 13: NATENVIY, National Spending on the Environment (2018)

	ORIGINAL	ADJUSTED	ABS DIFF
TOO LITTLE	67.4%	70.0%	(2.5%)
ABOUT RIGHT	24.3%	22.7%	(1.6%)
TOO MUCH	8.3%	7.3%	(0.9%)
AVERAGE DIFF			**1.7%**

Exhibit 14: NATEDUC, National Spending on Education (2018)

	ORIGINAL	ADJUSTED	ABS DIFF
TOO LITTLE	76.2%	75.2%	(1.0%)
ABOUT RIGHT	18.8%	18.9%	(0.1%)
TOO MUCH	5.0%	5.9%	(0.9%)
AVERAGE DIFF			**0.7%**

POSSLQ, relationship and cohabitation status, was asked on form X of ballots A, B, and C in 2018 and has statistically equivalent estimates. As shown in Exhibit 15, the differences in estimates across all categories are minimal, at less than 1 percentage point. The largest differences are seen in the "partner, living together" and "no spouse/partner" categories. The complement of this question, POSSLQY — which was asked on form Y of ballots A, B, and C — showed even smaller differences (ranging from 0.3 to 0.5 percentage points).

Exhibit 15: POSSLQ, Marital and Cohabitation Status (2018)

	ORIGINAL	ADJUSTED	ABS DIFF
MARRIED, LIVING TOGETHER	49.0%	49.0%	(0.0%)
PARTNER, LIVING TOGETHER	10.5%	11.3%	(0.7%)
MARRIED/PARTNERED, LIVING APART	5.3%	5.4%	(0.1%)
NO SPOUSE/PARTNER	35.2%	34.3%	(0.8%)
AVERAGE DIFF			**0.4%**

Another example, BIGBANG, which gauges respondents' knowledge of the beginning of the universe, was asked only on form X of ballot A in 2018, has statistically equivalent estimates and an average estimated difference. There is only a 0.9 percentage point difference in proportion saying "true" between the original and adjusted estimates, as shown in Exhibit 16, and this difference is not statistically meaningful. Our final example DISCAFFW — the likelihood that a qualified woman will be passed over for a promotion — was fielded on form Y of ballots A and B (the results were shown in Exhibit 5 earlier). The differences for this question are even smaller, 0.4 percentage points overall, and the estimates for individual categories are not meaningfully different.

Exhibit 16: BIGBANG, The Universe Began with a Huge Explosion (2018)

	ORIGINAL	ADJUSTED	ABS DIFF
TRUE	51.5%	52.4%	(0.9%)

Gender of the second person in the household (GENDER2) is another variable with a relatively large and potentially meaningful difference (1.9 points). Exhibit 17 presents the original and adjusted estimates and the absolute difference for the proportion of females in 2018. The proportion of females was slightly overestimated under the original weight (while the proportion of males was slightly underestimated). With both the original and adjusted estimates, females are still much more likely to be the second person in the household. These differences would be expected to change even though this question is included on all six ballot and form combinations, given that the assignment error resulted in ballot and form assignment inadvertently being related to the order of adults in the household.

Exhibit 17: GENDER2, Gender of the Second Person in the Household (2018)

	ORIGINAL	ADJUSTED	ABS DIFF
FEMALE	61.1%	59.1%	(1.9%)

Multivariate Analyses: The above examples indicate that overall topline conclusions appear unaffected by the assignment error. However, users may be concerned that the assignment error may reduce or nullify significant effects from multivariate analyses performed with the original estimates and weights. While we cannot rule out the possibility that users' multivariate analyses may be affected, our example analyses suggest that any differences should be minimal. In this section, we examine four commonly used measures from 2018: national spending on the environment (NATENVIY, from the Current Affairs topic); support for the death penalty (CAPPUN, from the Civil Liberties topic); whether "a working mother can establish just as warm and secure a relationship with her children as a mother who does not work" (FECHLD, from the Gender and Marriage topic); and whether the respondent has a gun at home (OWNGUN, from the Behaviors topic). These examples, like those above, represent various ballot and form combinations. NATENVIY was asked on form Y of all three ballots, while CAPPUN was asked on all ballots and forms. FECHLD and OWNGUN appeared on AX, AY, BX, and BY, and AX, AY, CX, and CY, respectively. NATENVIY has a potentially meaningfully large difference in 2018, while the other three variables have statistically equivalent estimates.

For each dependent variable, we fit two sets of models, first with the original weight and then with the adjusted weight. We include several controls in each model, including gender, age, marital status, race, education level, family income, and political ideology. In the first specification, we exclude those variables that are associated with the assignment error and are used during weight creation — gender, age, and marital status. We then fit a second model that includes these variables since many researchers will be interested in examining the effects of these directly or controlling for them in their models.

We present our results in Exhibits 18–21.[12] Looking across these examples, the direction and magnitude of the significant effects generally remain unchanged, while the different model specifications and weights do have some effect on our estimates. For example, political ideology is strongly positively and significantly associated with NATENVIY. This indicates that as political ideology becomes more conservative, the probability of choosing response options 2 ("about right") or 3 ("too much") increases. The magnitude and significance of this relationship is similar across the models tested.

For CAPPUN, we find that, across all specifications, Black and college-educated individuals are less likely to support the death penalty, while support increases with political conservatism. Our FECHLD models indicate that women, the college-educated, and those with higher incomes are more likely to believe that work does not affect a woman's relationship with her children. Finally, those who are white (as opposed to Black or other) and of higher income are more likely to own a gun.

Exhibit 18: Ordered Logistic Regression, Original and Adjusted Weights (NATENVIY, 2018)

	NATIONAL SPENDING ON THE ENVIRONMENT			
	ORIGINAL	ADJUSTED	ORIGINAL	ADJUSTED
BA DEGREE	-0.045	-0.067	-0.051	-0.070
SE	*0.176*	*0.186*	*0.176*	*0.183*
FAMILY INCOME	0.037	0.003	0.009	-0.041
SE	*0.079*	*0.085*	*0.089*	*0.090*
BLACK	-0.573	-0.461	-0.446	-0.365
SE	*0.289*	*0.296*	*0.298*	*0.311*
OTHER RACE	0.627*	0.475	0.733*	0.570*
SE	*0.287*	*0.291*	*0.280*	*0.286*
POL. IDEOLOGY	0.589***	0.582***	0.561***	0.550***
SE	*0.074*	*0.076*	*0.074*	*0.075*
FEMALE			-0.451***	-0.372*
SE			*0.127*	*0.146*
AGE			0.016***	0.016**
SE			*0.004*	*0.005*

[12] The negative sign on these coefficients indicates a lower probability of choosing response options 2 (agree), 3 (disagree), or 4 (strongly disagree).

	NATIONAL SPENDING ON THE ENVIRONMENT			
	ORIGINAL	ADJUSTED	ORIGINAL	ADJUSTED
PREV. MARRIED			-0.117	-0.086
SE			*0.187*	*0.181*
NEVER MARRIED			0.252	-0.201
SE			*0.253*	*0.293*
THRESHOLD 1	3.199^{***}	3.617^{***}	3.262^{***}	3.644^{***}
SE	*0.350*	*0.431*	*0.363*	*0.458*
THRESHOLD 2	5.021^{***}	5.481^{***}	5.077^{***}	5.498^{***}
SE	*0.377*	*0.440*	*0.402*	*0.483*
N	**1,010**	**1,010**	**1,007**	**1,007**

Note: Dependent variable is NATENVIY. Response options range from "too little" (1) to "too much" (3). Models 1 and 3 exclude variables used in weighting. Standard errors in italics, reference categories include male; married, white, and some college or less. Family income (CONINC) was standardized. Political ideology ranges from extremely liberal (1) to extremely conservative (7). * *p < 0.05,* ** *p < 0.01,* *** *p < 0.001*

Exhibit 19: Logistic Regression, Original and Adjusted Weights (CAPPUN, 2018)

	SUPPORT FOR THE DEATH PENALTY			
	ORIGINAL	ADJUSTED	ORIGINAL	ADJUSTED
BA DEGREE	-0.631^{***}	-0.537^{***}	-0.621^{***}	-0.527^{**}
SE	*0.123*	*0.124*	*0.124*	*0.123*
FAMILY INCOME	0.043	0.001	0.036	-0.018
SE	*0.066*	*0.073*	*0.070*	*0.077*
BLACK	-0.950^{***}	-0.964^{***}	-0.968^{***}	-0.989^{***}
SE	*0.156*	*0.166*	*0.159*	*0.167*
OTHER RACE	-0.348	-0.324	-0.369	-0.352
SE	*0.214*	*0.213*	*0.222*	*0.221*
POL. IDEOLOGY	0.361^{***}	0.390^{***}	0.364^{***}	0.395^{***}
SE	*0.040*	*0.043*	*0.041*	*0.044*
FEMALE			-0.253^{*}	-0.210
SE			*0.127*	*0.134*
AGE			-0.003	-0.004
SE			*0.004*	*0.005*
PREV. MARRIED			0.087	0.096
SE			*0.163*	*0.174*
NEVER MARRIED			-0.025	0.004
SE			*0.152*	*0.161*
CONSTANT	-0.487^{**}	-0.598^{***}	-0.214	-0.345
SE	*0.161*	*0.171*	*0.260*	*0.291*
N	**1,975**	**1,975**	**1,970**	**1,970**

Note: Dependent variable is CAPPUN. 1 (0) indicates support (opposition). Models 1 and 3 exclude variables used in weighting. Standard errors in italics, reference categories include male; married, white, and some college or less. Family income (CONINC) was standardized. Political ideology ranges from extremely liberal (1) to extremely conservative (7). * *p < 0.05,* ** *p < 0.01,* *** *p < 0.001*

Exhibit 20: Ordered Logistic Regression, Original and Adjusted Weights (FECHLD, 2018)

	WORKING MOTHER CAN HAVE WARM RELATIONSHIP WITH CHILD			
	ORIGINAL	ADJUSTED	ORIGINAL	ADJUSTED
BA DEGREE	-0.525***	-0.440*	-0.568***	-0.474**
SE	*0.135*	*0.167*	*0.134*	*0.170*
FAMILY INCOME	-0.166*	-0.179*	-0.191**	-0.201*
SE	*0.071*	*0.077*	*0.069*	*0.078*
BLACK	0.046	0.036	0.062	0.022
SE	*0.156*	*0.153*	*0.161*	*0.159*
OTHER RACE	0.447*	0.456*	0.542**	0.509*
SE	*0.185*	*0.193*	*0.192*	*0.200*
POL. IDEOLOGY	0.225***	0.241***	0.202***	0.218***
SE	*0.040*	*0.046*	*0.038*	*0.043*
FEMALE			-0.614***	-0.667***
SE			*0.133*	*0.147*
AGE			0.017***	0.014**
SE			*0.004*	*0.004*
PREV. MARRIED			-0.147	-0.050
SE			*0.175*	*0.195*
NEVER MARRIED			0.072	0.097
SE			*0.166*	*0.175*
THRESHOLD 1	0.059	0.159	0.401	0.336
SE	*0.186*	*0.207*	*0.292*	*0.300*
THRESHOLD 2	1.952***	2.024***	2.366***	2.270***
SE	*0.210*	*0.237*	*0.327*	*0.342*
THRESHOLD 3	4.002***	3.986***	4.453***	4.267***
SE	*0.221*	*0.246*	*0.321*	*0.342*
N	**1,379**	**1,379**	**1,375**	**1,375**

Note: Dependent variable is FECHLD. Response options range from "strongly agree" (1) to "strongly disagree" (4). Models 1 and 3 exclude variables used in weighting. Standard errors in italics, reference categories include male; married, white, and some college or less. Family income (CONINC) was standardized. Political ideology ranges from extremely liberal (1) to extremely conservative (7). * p < 0.05, ** p < 0.01, *** p < 0.001

Exhibit 21: Logistic Regression, Original and Adjusted Weights (OWNGUN, 2018)

	ORIGINAL	ADJUSTED	ORIGINAL	ADJUSTED
BA DEGREE	-0.192	-0.204	-0.283	-0.299
SE	*0.168*	*0.169*	*0.178*	*0.171*
FAMILY INCOME	0.346***	0.369***	0.261***	0.286***
SE	*0.072*	*0.078*	*0.073*	*0.081*
BLACK	-0.688**	-0.544*	-0.559*	-0.407
SE	*0.243*	*0.251*	*0.246*	*0.250*
OTHER RACE	-1.122***	-1.018***	-1.275***	-1.026***
SE	*0.237*	*0.233*	*0.246*	*0.245*
POL. IDEOLOGY	0.312***	0.293***	0.293***	0.289***
SE	*0.049*	*0.049*	*0.049*	*0.050*
FEMALE			-0.452**	-0.500***
SE			*0.136*	*0.146*
AGE			-0.001	0.000
SE			*0.004*	*0.004*
PREV. MARRIED			-0.575**	-0.436*
SE			*0.181*	*0.189*
NEVER MARRIED			-0.877***	-0.870***
SE			*0.187*	*0.185*
CONSTANT	-1.547***	-1.617***	-0.768*	-0.918**
SE	*0.225*	*0.230*	*0.337*	*0.318*
N	**1,365**	**1,365**	**1,363**	**1,363**

Note: Dependent variable is OWNGUN. 1 indicates respondent has a gun at home. Models 1 and 3 exclude variables used in weighting. Standard errors in italics, reference categories include male; married, white, and some college or less. Family income (CONINC) was standardized. Political ideology ranges from extremely liberal (1) to extremely conservative (7). * p < 0.05, ** p < 0.01, *** p < 0.001

Discussion and Conclusion

This report summarizes our final analysis of the assignment error that occurred in how respondents were assigned to the GSS questionnaire ballots and forms in 2002, 2010, 2012, 2016, and 2018. NORC created weights to adjust for this error called BALLOTFORMWT and BALLOTFORMWTNR. These are individual-level rebalanced weights of WTSSALL for affected years and will be released with Release 3 of the 2021 GSS cumulative data file. These new weights should be used in most cases when analyzing questions that are asked on only some ballots or forms in the affected years. In addition, Release 3 includes the original unadjusted weights WTSS, WTSSALL, and WTSSNR, so interested researchers can conduct their own investigations into the impact of the error on their specific analyses. Release 3 of the GSS 2021 cumulative file also includes the new post-stratified weights extending back to 2000: WTSSPS (2000–2021) and WTSSNRPS (2004–2021).

Our analysis comparing the estimates for all questions in the affected years using the original and corrected weights indicates that overall impacts are small — with the average difference across categories less than 1 percentage point. Furthermore, only 24 questions (1.2 percent) asked on only some forms or ballots across the affected years were potentially meaningful differences based on our statistical equivalence testing, and even fewer questions (4 questions) had a difference of 3 percentage points or greater.

The error is mitigated for most questions because they are asked on multiple ballots and forms, and thus the impact of the error was partially or entirely offset. The question wording experiments that use form had the most potential to be impacted by the assignment error, given the number of households impacted. Nonetheless, the analysis shows that the impacts are small for most questions asked on only one form, although a few questions do show larger impacts. Overall, our results indicate that nearly all substantive conclusions drawn from analysis of GSS data based on simple associations will be the same whether analysts use the original or the new weights. While we cannot rule out the possibility that users' multivariate analyses may be affected, our example analyses suggest that any differences should be minimal.

Methodology

Rebalanced Weights: We produced rebalanced versions of WTSSALL and WTSSNR. The analysis in this report evaluated the original and adjusted WTSSALL weight in order to focus on the impact of the assignment error and ability of weights to adjust for it, separate from nonresponse adjustments.

The procedure to rebalance the weights comprised the following steps:

1. Identify the years of data collection affected by the assignment error. We determined these to be 2002, 2010, 2012, 2016, and 2018.
2. Determine the combinations of ballot and form where a rebalancing is needed. To accomplish the most complete and flexible solution, such that any GSS question regardless of which ballot(s) and form(s) on which it appeared could be properly analyzed, the weighting corrections were applied separately for each of the six combinations of ballot (A, B, C) and form (X, Y) in each affected year. While this approach carried the most potential for increased standard errors among items that appear in multiple combinations, results showed these increases were very small.
3. Isolate the households affected by the assignment error. The assignment error by ballot affected households with multiples of 3 adults (3-adult and 6-or-more-adult households), while the assignment error by form affected households with multiples of 2 adults (2-adult, 4-adult, and 6-or-more-adult households). As a result, we considered any respondents in 2-adult, 3-adult, 4-adult, or 6-or-more-adult households affected.
4. Identify the dimensions on which to rebalance the weights. A preliminary analysis found that four demographic variables were associated with ballot and form assignment and had become skewed because of the assignment error. These variables were age, sex marital status, and work status.

5. Choose a starting weight. WTSSALL and WTSSNR, the two final weights recommended for data users, were each rebalanced following this procedure.
6. Impute missing values in the raking dimensions. Each of age, marital status, and work status contained a small number of missing values due to item nonresponse or “don’t know” and “refused” responses. Missing values were imputed, separately for each year, using the hot-deck method. Sex had no missing values and was not imputed.
7. Compute weighted totals for the raking dimensions across the full sample of affected households from each year, using the original GSS weight. The following groupings were used to produce the totals for the raking procedure:
    - **AGE:** 5-level age group
        - 18–29
        - 30–39
        - 40–49
        - 50–64
        - 65+
    - **SEX**: 2-level
        - Male
        - Female
    - **MARITAL:** 3-level marital status
        - Never married
        - Married
        - Post-married
    - **WRKSTAT:** 3-level work status
        - Full-time employed
        - Part-time, temporarily not working, or unemployed
        - Not in labor force
8. Within each combination of ballot and form, and among affected households, rake weights to align with the cross-ballot/form totals among all affected households. The raking procedure begins with the first dimension (AGE) and applies ratio adjustments to the original weights within each level (e.g., the 5 age groups) such that the resulting adjusted weights sum to the cross-ballot/form totals. The process then moves to the next dimension (SEX), and ratio adjusts in the same manner. After these adjustments have been made for all four dimensions, the process starts over again with dimension one. It works circularly through the dimensions, applying ratio adjustments until the new adjusted weights agree with the pre-established totals across all four dimensions simultaneously. At the end of this process, the newly corrected weights will agree, within each combination of ballot and form, with the totals that were produced across all combinations of ballot and form among affected households. As a result, each ballot/form subset is representative of the full sample of respondents, which corrects the skews caused by the randomization issue.

Estimated design effects under both the original and adjusted WTSSALL were calculated as $DEFF = 1 + CV^2$, where CV is the coefficient of variation of the weights. Simulated margins of error were calculated as $1.96 \times \sqrt{\frac{0.25 \times DEFF}{n}}$. This represents an anticipated 95 percent confidence interval half-width around a hypothetical survey estimate of 50 percent.

Analysis: Variables from affected years (2002, 2010, 2012, 2016, and 2018) were included for analysis if they had greater than 25 non-missing observations. The "don't know," "refused," and "skipped on web" response categories were treated as missing after determining that the distribution of these responses was similar under both the original and corrected weights (average percentage point difference of 0.06 (range = 0.0-0.23) across years).

The difference analysis examined discrete variables (dichotomous, ordered, and unordered). We estimated the weighted proportion of each response category for each variable using both the original and new weights. This produced two sets of category-level estimates. The adjusted estimates were then subtracted from the original estimates, and the absolute value was taken. The final variable-level estimate was produced by averaging over the category-level difference estimates for each variable. For continuous variables, we analyzed the average absolute difference in mean estimates across original and adjusted weights. A summary of the continuous difference results can be found in Appendix D.

Equivalence testing was also undertaken on both discrete and continuous variables. Discrete variables were tested at the category level, and categories were excluded from testing if they contained 25 or fewer observations. Dichotomous variables were created for each included category. Continuous variables were not transformed.

Statistical equivalence testing using the TOST (two one-sided test) procedure starts by defining a smallest effect size of interest, or Δ. We defined Δ as the margin of error of the category mean (in the case of discrete variables) or variable mean (in the case of continuous variables) under the original weights. After obtaining the margin of error, the dataset containing the original weight was appended to one containing the adjusted weight, resulting in a stacked dataset with two observations per respondent (denoted by an "id" variable).

Next, we estimated weighted regression models with standard errors clustered by respondent for each variable in the dataset, with each respective variable serving as the outcome and a binary indicator for "corrected weight" as the sole predictor. The coefficient on this predictor represents the difference between the mean estimates under the original and adjusted weights. Each difference was then subjected to two one-sided hypothesis tests:

H01: (μ1 - μ2) >= Δ versus Ha1: (μ1 - μ2) < Δ
H02: (μ1 - μ2) <= -Δ versus Ha1: (μ1 - μ2) > -Δ

The weighted mean estimates are declared statistically equivalent if both of the above null hypotheses are rejected at the chosen α level of 5%.

References

General Social Surveys 1972-2018 Cumulative Codebook. 2019.
https://gss.norc.org/documents/codebook/gss_codebook.pdf

Kish, Leslie. (1949). A procedure for objective respondent selection within a household. *Journal of the American Sociological Association*, 44, 380-387.

Lakens, Daniel, Anne M. Scheel, and Peder M. Isager. (2018). "Equivalence testing for psychological research: A tutorial." *Advances in Methods and Practices in Psychological Science* 1, no. 2: 259-269.

Smith, Tom W., and Peterson, Bruce L. (1986). “Problems in form randomization on the General Social Surveys.” *GSS Methodological Report* No. 36.
https://gss.norc.org/Documents/reports/methodological-reports/MR036.pdf

Appendix

Appendix A: Assignment Error by Ballot (A, B, C) and by Form (X, Y) in 2018

# of Adults		Household Age Rank (Oldest to Youngest)					
	Ballot	**1**	**2**	**3**	**4**	**5**	**6**
1	**A**	274					
	B	253					
	C	248					
2	**A**	185	198				
	B	199	196				
	C	204	214				
3	**A**	.	87	.			
	B	89	.	.			
	C	.	.	78			
4	**A**	7	9	5	14		
	B	6	5	5	16		
	C	5	9	10	15		
5	**A**	1	.	1	1	1	
	B	.	.	1	.	1	
	C	3	.	1	2	.	
6	**A**	2	.	.	.	.	.
	B	.	.	2	.	.	1
	C	.	.	.	.	.	.
	Form	**1**	**2**	**3**	**4**	**5**	**6**
1	**X**	380					
	Y	395					
2	**X**	.	608				
	Y	588	.				
3	**X**	42	47	37			
	Y	47	40	41			
4	**X**	.	23	.	45		
	Y	18	.	20	.		
5	**X**	1	.	2	1	1	
	Y	3	.	1	2	1	
6	**X**	.	.	.	.	.	1
	Y	2	.	2	.	.	.

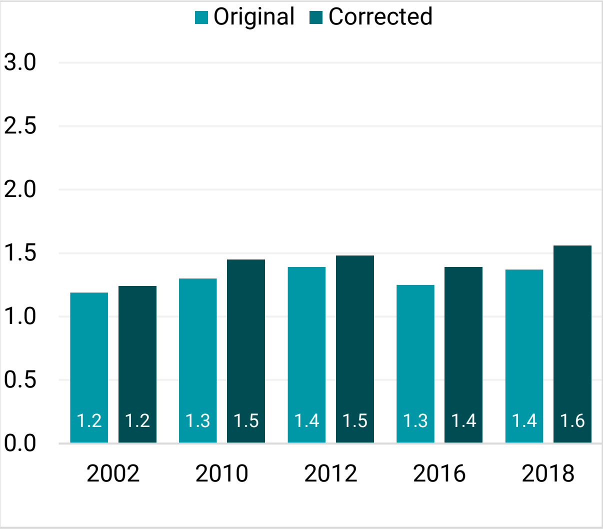
Note: Count of respondents assigned to each ballot and form

Appendix B: Design Effects and Margins of Error

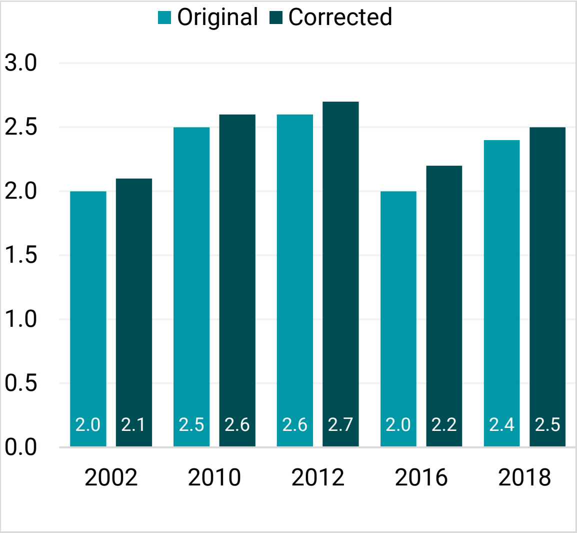
In general, the new weights produce modest increases in estimated design effects and simulated margins of error for hypothetical estimates. In other words, using the adjusted weights will result in minimal loss of precision in estimation. This was found to be true at both the individual ballot/form level as well as across ballots and forms and for the full sample in each year.

Below we present the estimated design effects under the original values compared to the adjusted values for WTSSALL for each of the affected years. The comparison shows a maximum of only two-tenths increase in the estimated design effects for any year. We also present the simulated margins of error on a hypothetical 50-percent estimate for each affected year, again comparing the original and adjusted WTSSALL, and the results similarly show only a one-tenth or two-tenths increase in anticipated margins of error.

Estimated Design Effects Using Original and Corrected Weights by Year

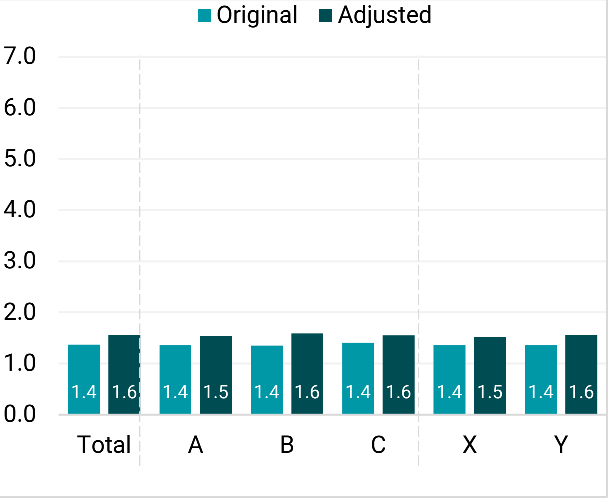


Simulated Margins of Error for a 50-Percent Estimate Using Original and Corrected Weights by Year

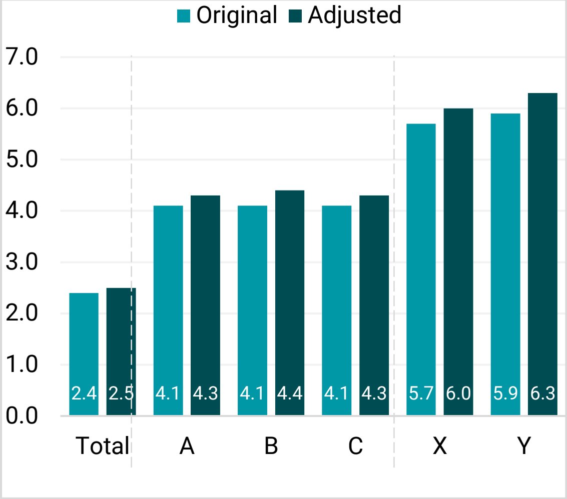


The figures below show estimated design effects and simulated margins of error, respectively, by ballot and form, using the affected year 2018 as an example. While the anticipated margins of error may increase by a slightly larger amount when looking at the specific ballot and form samples (up to four-tenths of a percent), the loss of precision from the rebalanced weights remains small.

Estimated Design Effects Using Original and Adjusted Weights, Overall and by Ballot and Form (2018)



Simulated Margins of Error for a 50-Percent Estimate Using Original and Adjusted Weights, Overall and by Ballot and Form (2018)



The table below shows the coefficients of variation, estimated design effects, and simulated margins of error for the full sample and for each ballot and form by year for the affected years.

Original and Adjusted Coefficients of Variation (CV), Estimated Design Effects (DEFF), and Simulated Margins of Error (MOE) by Year

	TOTAL		BALLOT A		BALLOT B		BALLOT C		FORM X		FORM Y	
CV	ORIG.	ADJ.	ORIG.	ADJ.	ORIG.	ADJ.	ORIG.	ADJ.	ORIG.	ADJ.	ORIG.	ADJ.
2002	43.4	48.6	43.1	47.5	43.0	51.1	44.1	47.3	44.0	50.6	42.0	44.1
2010	54.4	67.0	55.8	66.2	54.2	67.8	53.1	67.0	55.0	63.1	56.6	69.2
2012	62.3	69.5	65.2	67.0	60.7	70.0	61.1	71.4	59.9	64.2	70.4	69.9
2016	49.8	62.7	51.2	66.6	47.5	58.6	50.5	62.8	53.8	59.7	47.3	73.7
2018	61.1	74.6	60.1	73.4	58.9	76.5	64.0	73.9	59.9	72.2	60.1	74.6
DEFF	ORIG.	ADJ.	ORIG.	ADJ.	ORIG.	ADJ.	ORIG.	ADJ.	ORIG.	ADJ.	ORIG.	ADJ.
2002	1.2	1.2	1.2	1.2	1.2	1.3	1.2	1.2	1.2	1.3	1.2	1.2
2010	1.3	1.5	1.3	1.4	1.3	1.5	1.3	1.5	1.3	1.4	1.3	1.5
2012	1.4	1.5	1.4	1.5	1.4	1.5	1.4	1.5	1.4	1.4	1.5	1.5
2016	1.3	1.4	1.3	1.4	1.2	1.3	1.3	1.4	1.3	1.4	1.2	1.5
2018	1.4	1.6	1.4	1.5	1.4	1.6	1.4	1.6	1.4	1.5	1.4	1.6
MOE	ORIG.	ADJ.	ORIG.	ADJ.	ORIG.	ADJ.	ORIG.	ADJ.	ORIG.	ADJ.	ORIG.	ADJ.
2002	2.0	2.1	3.5	3.6	3.5	3.6	3.5	3.6	5.0	5.2	4.9	4.9
2010	2.5	2.6	4.3	4.6	4.0	4.3	4.5	4.8	6.1	6.3	6.2	6.6
2012	2.6	2.7	4.6	.47	4.4	4.6	4.4	4.6	6.3	6.5	6.8	6.8
2016	2.0	2.2	3.6	3.9	3.5	3.6	3.5	3.7	5.2	5.3	5.1	5.7
2018	2.4	2.5	4.1	4.3	4.1	4.4	4.1	4.3	5.7	6.0	5.9	6.3

APPENDIX C: DEMOGRAPHIC TABLES FOR ALL YEARS

Original and Adjusted Proportion 18–49 and 50+

	TOTAL		BALLOT A		BALLOT B		BALLOT C		FORM X		FORM Y	
% 18–49	ORIG.	ADJ.	ORIG.	ADJ.	ORIG.	ADJ.	ORIG.	ADJ.	ORIG.	ADJ.	ORIG.	ADJ.
2002	62.1%	62.1%	64.9%	61.9%	60.4%	62.3%	61.0%	62.2%	63.7%	61.4%	60.6%	62.8%
2010	56.6%	56.6%	56.7%	55.8%	57.2%	56.8%	55.9%	57.3%	53.3%	55.6%	60.0%	57.7%
2012	58.4%	58.4%	62.6%	60.1%	59.1%	58.4%	53.6%	56.8%	63.0%	58.7%	53.6%	58.1%
2016	53.5%	53.5%	49.9%	54.0%	57.6%	52.8%	52.8%	53.6%	57.2%	53.6%	49.7%	53.3%
2018	56.6%	56.6%	55.5%	56.6%	50.2%	56.5%	63.6%	56.5%	61.8%	56.5%	50.8%	56.6%
% 50+	ORIG.	ADJ.	ORIG.	ADJ.	ORIG.	ADJ.	ORIG.	ADJ.	ORIG.	ADJ.	ORIG.	ADJ.
2002	37.9%	37.9%	35.1%	38.1%	39.6%	37.7%	39.0%	37.8%	36.3%	38.6%	39.4%	37.2%
2010	43.4%	43.4%	43.3%	44.2%	42.8%	43.2%	44.1%	42.7%	46.7%	44.4%	40.0%	42.3%
2012	41.6%	41.6%	37.4%	39.9%	40.9%	41.6%	46.4%	43.2%	37.0%	41.3%	46.4%	41.9%
2016	46.5%	46.5%	50.1%	46.0%	42.4%	47.2%	47.2%	46.4%	42.8%	46.4%	50.3%	46.7%
2018	43.4%	43.4%	44.5%	43.4%	49.8%	43.5%	36.4%	43.5%	38.2%	43.5%	49.2%	43.4%

Original and Adjusted Proportion Female and Male

	TOTAL		BALLOT A		BALLOT B		BALLOT C		FORM X		FORM Y	
% FEMALE	ORIG.	ADJ.	ORIG.	ADJ.	ORIG.	ADJ.	ORIG.	ADJ.	ORIG.	ADJ.	ORIG.	ADJ.
2002	54.2%	54.2%	55.1%	54.2%	54.2%	55.0%	53.4%	53.4%	54.5%	54.0%	53.9%	54.4%
2010	54.8%	54.8%	55.0%	53.5%	57.4%	56.8%	51.5%	53.9%	46.8%	55.1%	62.9%	54.6%
2012	53.9%	53.9%	53.3%	54.0%	53.0%	54.4%	55.4%	53.4%	59.6%	54.1%	48.0%	53.7%
2016	54.8%	54.8%	54.9%	55.3%	53.5%	54.5%	55.9%	54.5%	62.3%	54.7%	47.1%	54.9%
2018	54.5%	54.5%	55.7%	54.2%	53.5%	54.6%	54.2%	54.6%	61.3%	53.8%	47.0%	55.1%
% MALE	ORIG.	ADJ.	ORIG.	ADJ.	ORIG.	ADJ.	ORIG.	ADJ.	ORIG.	ADJ.	ORIG.	ADJ.
2002	45.8%	45.8%	44.9%	45.8%	45.8%	45.0%	46.6%	46.6%	45.5%	46.0%	46.1%	45.6%
2010	45.2%	45.2%	45.0%	46.5%	42.6%	43.2%	48.5%	46.1%	53.2%	44.9%	37.1%	45.4%
2012	46.1%	46.1%	46.7%	46.0%	47.0%	45.6%	44.6%	46.6%	40.4%	45.9%	52.0%	46.3%
2016	45.2%	45.2%	45.1%	44.7%	46.5%	45.5%	44.1%	45.5%	37.7%	45.3%	52.9%	45.1%
2018	45.5%	45.5%	44.3%	45.8%	46.5%	45.4%	45.8%	45.4%	38.7%	46.2%	53.0%	44.9%

Appendix D: Full Analysis Results

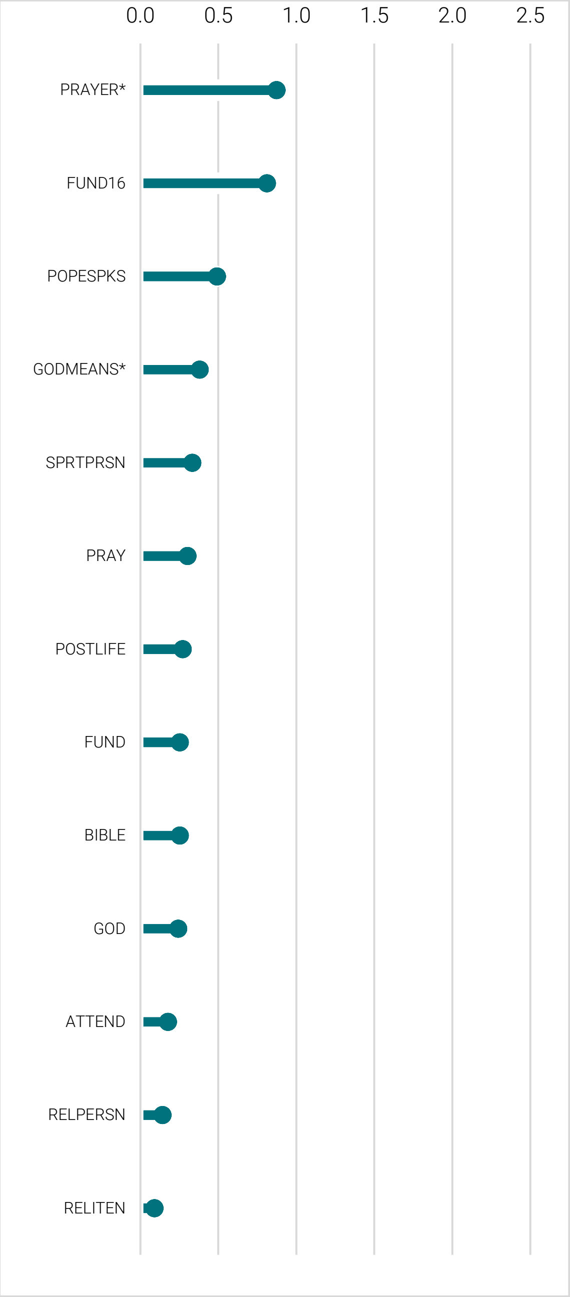
This spreadsheet (https://gss.norc.org/Documents/reports/methodological-reports/MR134%20supplement.zip) has a variable-level and category-level tab for each of the affected years.

- The variable-level sheets include the variable name, unweighted N, type of variable, average difference in proportions (discrete) or difference in means (continuous) and average % change across weighted estimates, indicates which ballots and forms the question was asked on, and whether the question had potentially meaningful differences based on the statistical equivalence testing.
- The category-level includes the equivalence testing details for the questions with potentially meaningful differences. It shows the category, unweighted N for the question and for that category, the original and corrected mean, the original and corrected standard error, the margin of error for the original estimate for that category, the estimated difference for that category and the standard error of the estimated difference, as well as the upper and lower bounds of the t-values and p-values.

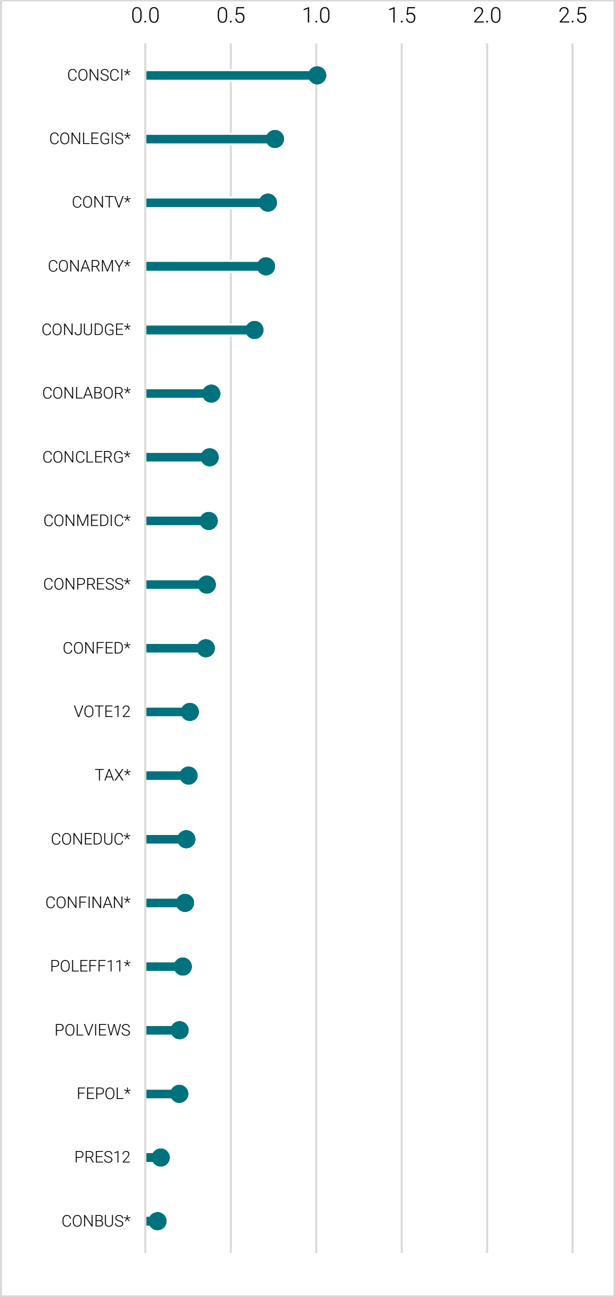
Summary of Difference in Means for Continuous Variables

	NUMBER OF VARIABLES	MEDIAN DIFFERENCE ACROSS CATEGORIES
2018	132	0.10
2016	102	0.12
2012	90	0.14
2010	82	0.07
2002	116	0.05
TOTAL	522	0.09

Avg. Percentage Point Difference, Religion and Spirituality (2018)

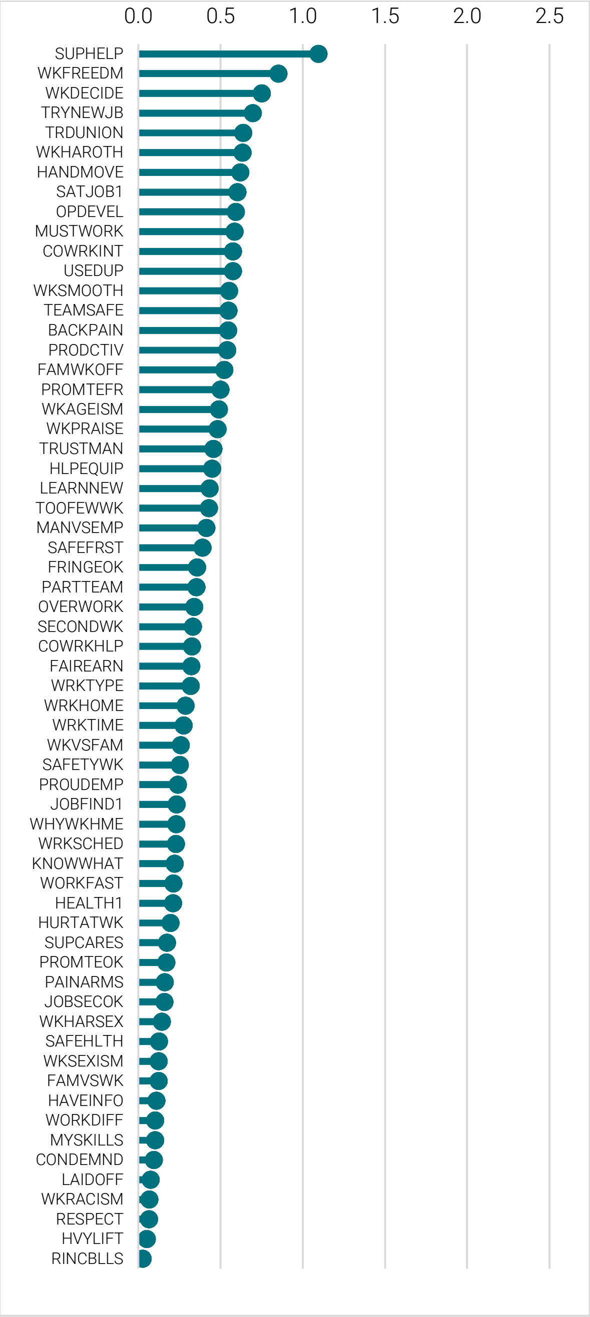


Avg. Percentage Point Difference, Politics (2018)

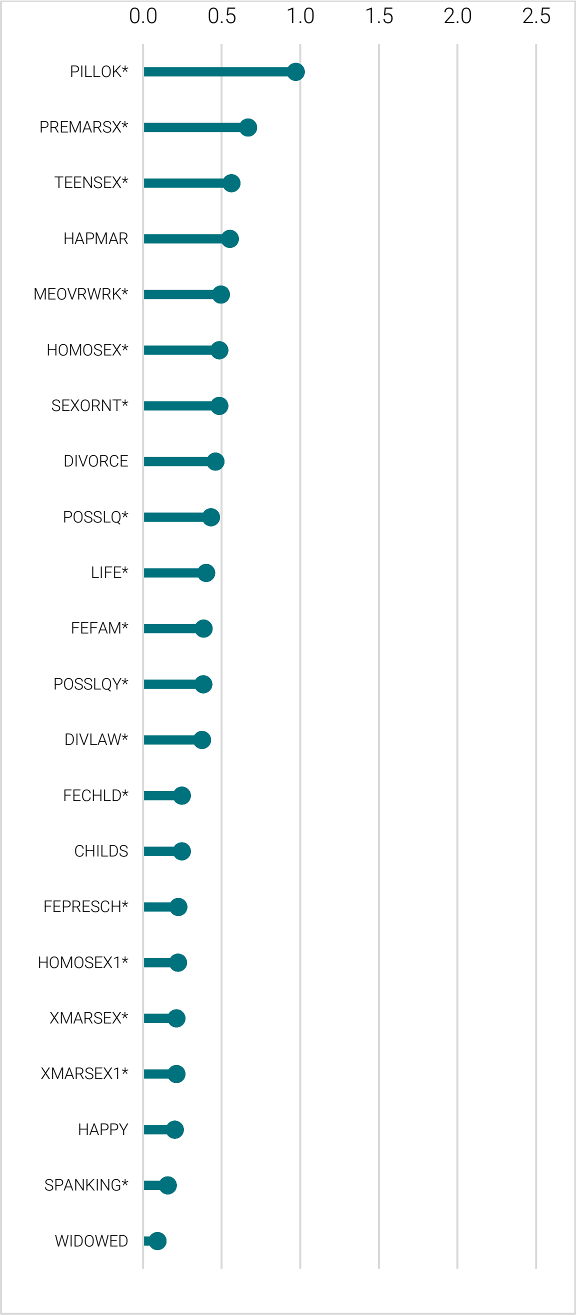


Note: Asterisks (*) in the above figures indicate questions that were asked only on some forms or ballots.

Avg. Percentage Point Difference, Quality of Working Life (2018)



Avg. Percentage Point Difference, Gender and Marriage (2018)



Note: Asterisks (*) in the above figures indicate questions that were asked only on some forms or ballots.