Title

Reliability of the core items in the General Social Survey: Estimates from the three-wave panels, 2006-2014

Authors
Hout, M
Hastings, OP

Publication Date
2016-11-14

DOI
10.15195/v3.a43

sociological science

Reliability of the Core Items in the General Social Survey: Estimates from the Three-Wave Panels, 2006–2014

Michael Hout,[a] Orestes P. Hastings[b]

a) New York University; b) University of California, Berkeley

Abstract: We used standard and multilevel models to assess the reliability of core items in the General Social Survey panel studies spanning 2006 to 2014. Most of the 293 core items scored well on the measure of reliability: 62 items (21 percent) had reliability measures greater than 0.85; another 71 (24 percent) had reliability measures between 0.70 and 0.85. Objective items, especially facts about demography and religion, were generally more reliable than subjective items. The economic recession of 2007–2009, the slow recovery afterward, and the election of Barack Obama in 2008 altered the social context in ways that may look like unreliability of items. For example, unemployment status, hours worked, and weeks worked have lower reliability than most work-related items, reflecting the consequences of the recession on the facts of peoples lives. Items regarding racial and gender discrimination and racial stereotypes scored as particularly unreliable, accounting for most of the 15 items with reliability coefficients less than 0.40. Our results allow scholars to more easily take measurement reliability into consideration in their own research, while also highlighting the limitations of these approaches.

Keywords: reliability; survey methodology; General Social Survey; measurement error

Citation: Hout, Michael, and Orestes P. Hastings. 2016. "Reliability of the Core Items in the General Social Survey: Estimates from the Three-Wave Panels, 2006–2014." Sociological Science 3: 971-1002.

Received: August 27, 2016

Accepted: October 4, 2016

Published: November 14, 2016

Editor(s): Jesper Sørensen, Gabriel Rossman

DOI: 10.15195/v3.a43

SURVEY research works on a very simple "we ask; they tell" principle. Researchers approach a representative sample of individuals, ask some questions, record the answers, and interpret the results (usually with the help of descriptive and inferential statistics). If people cannot or will not answer the questions (or if they answer them incorrectly or untruthfully), the whole enterprise, and the social science grounded in it, suffers.

To assess the quality of people's reports about themselves, researchers usually repeat the questions and compare all the answers. Questions are deemed to be reliable if people answer them more or less the same every time; a question is deemed to be unreliable if people have trouble answering consistently. Typically researchers wait awhile (sometimes years) before repeating a question. They do not want to annoy respondents, and they do not want memories of past answers to contaminate current reports, because the common statistical models of reliability assume that any errors in the answers are independent of one another. Sometimes researchers can assess reliability by comparing people's answers to data from other sources; voting records are public, and tax authorities have occasionally granted access to income records for research purposes.

Unreliability is but one component of total survey error (Alwin 2007:3–12). We focus on it here because we have a special resource for addressing it. Few surveys contain repeated measures; researchers working with those data have no choice but to treat the data as if it were perfectly reliable, even though no one believes it really is. They may temper their interpretations, but they have few statistical tools.

Two repeated measures of the same item can show difference, but the researcher cannot separate real change from response errors. Three measures offer the prospect, with generally reasonable assumptions, of measuring both change and reliability (Alwin 2007:95–148; Heise 1969; Wiley and Wiley 1970). We exploit three-wave panel data from the General Social Survey (GSS), starting in 2006 and ending in 2014. The GSS reinterviewed the respondents from 2006 in 2008 and 2010, reinterviewed the respondents from 2008 in 2010 and 2012, and reinterviewed the respondents from 2010 in 2012 and 2014. Thus, we have the unique opportunity to assess reliability for a broad set of widely used social variables.

The GSS is among the most-used data sets in the social sciences. The GSS offers a nearly unique combination of long time series, representative samples, and very broad content. It is the primary source of information for more than 200 "core" items: the facts, attitudes, values, and opinions that are always in the GSS. Basic demographic facts are available from the U.S. Census Bureau and elsewhere, but representative data on other variables in the GSS core are harder to find.

Assessing the quality of the GSS data is a crucial task. We know a great deal about the quality of the GSS sample (e.g., Smith et al. 2015, Appendix A). For example, in an era of low response rates, the five most recent GSSs had response rates greater than 70 percent. But we have known much less about the quality of the questions that are the substance of the survey. The first line of quality control is, of course, the selection of good questions. Some core items replicate questions asked in other surveys (e.g., Davis and Smith 1980, appendix N). All new GSS items are pretested and vetted through cognitive interviews before fieldwork begins. A second part of quality control is to check related items for consistency. Items that contribute to scales can be checked using item-to-item variation within the scale (e.g., Clogg and Sawyer 1981; Treiman 2009:243–257).

The highest standard of question reliability comes from repeated measures. If people answer the same question the same way again and again, then that question is, by definition, reliable. However, people also change in various ways over time, and a good measure of reliability will also account for that. The GSS panels that ran from 2006 to 2014 provide repeated measures that allow us the opportunity to assess the reliability of 283 core items and about 20 other variables derived from them. We use two complementary measures of reliability: Alwin's (2007) adaptation of the estimator proposed by Heise (1969) and a simple multilevel model of our own devising. Alwin (2007) used variations of the Heise approach to estimate reliabilities for many variables in the American National Election Study (ANES) panels of the 1950s, 1970s, and 1990s, plus three other three-wave panel studies. Hout and Hastings (2012) did the same for the 2006–2010 GSS panels.

Our goal is to make a broad-brush assessment of question quality across the entire questionnaire, as Alwin (2007) did for the ANES. Much more can be learned from detailed examination of a small number of items (e.g., Duncan, Stenbech, and Brody 1988). In particular, a model tailored to particular items and the substantive issues of interest can advance thinking in a specific domain. We trust that the research community will undertake this kind of close examination as the GSS panel data become better known. But as an initial foray into the quality of the GSS as a whole, especially its core items, it is important to exploit the panels to get basic estimates of the reliability of all the variables.

We presented a preliminary assessment in a methodological report for the GSS (Hout and Hastings 2012). That report used only the Heise (1969) method of estimating reliability, analyzed only the 1,238 respondents from the 2006–2010 panel, and used a crude scheme for grouping variables. All in all, it was aimed at a community of GSS users who were most interested in a few specific items and did little to address questions of broader sociological interest.

We present the reliability models in the next section. In subsequent sections, we discuss some decisions we had to make about specific questions; classify the variables into major types and subtypes; present the main results by major type, subtype, and single variable; and briefly analyze a few items that changed appreciably between 2006 and 2014. We conclude by proposing an agenda for future research on the quality of GSS variables.

Quantifying Reliability

Classical measurement theory (e.g., Alwin 2007) distinguishes between a measurement (y) taken as part of a survey (or in some other form of data collection) and its true value (Y); the difference being measurement error (ϵ). Total variance (σ_y^2) is the sum of true score variance (σ_Y^2), error variance (σ_ϵ^2), and double the covariance between the true score and the error ($\text{cov}(Y, \epsilon)$). Reliability (ρ) is defined as the ratio of the true score variance to the total variance. It is customary to think of measurement error as uncorrelated with true scores, in which case the covariance of Y and ϵ is zero and it drops out of the equation. The total variance is then just the sum of the true score variance and error variance, and assessing reliability becomes a matter of finding a way to separate the two.

A single measurement cannot separate total variance into true score variance and measurement error variance; it takes at least two measurements. Some concepts lend themselves to the use of multiple measures within one survey; for instance, the GSS measures vocabulary with a ten-word quiz. But essential variables such as religious denomination or political party can only be asked once in a given survey without respondents wondering why they are being asked to repeat their answer. For these variables, the only way to get repeated measures is to conduct a second interview some time later. This article is about using three measures taken at two-year intervals to estimate reliability, focused specifically on the GSS.

Two measures suffice if it is reasonable to assume that the true score Y has not changed between the first and second measurement. For constant Y, the difference between the first and second measurements (y_{i1} and y_{i2}) must be due to measurement error, and the covariance equals the true score variance, σ_Y^2. Furthermore, the product of the standard deviation of y_1 and the standard deviation of y_2 is an estimate of total variance ($\sigma_Y^2 + \sigma_\epsilon^2$). That implies an estimate of reliability:

$$\hat{\rho} = \frac{Cov(y_1, y_2)}{s_1 s_2} \tag{1}$$

where $s_t(t = 1, 2)$ indicates the standard deviation of the measures at each time. Ideally, the two measurements are separated by enough time to ensure that few respondents recall their first answer when they are being reinterviewed but close enough in time that assuming Y did not change is plausible. Of course, some true scores do not change by definition. For example, the GSS has many questions about when a person was growing up. Those should not change no matter how much time passed between interviews; Smith and Son (2011) analyzed the stability of those items between 2006 and 2008.

Measuring a variable a third time offers substantially more statistical power for assessing reliability. It makes it possible for us to estimate reliability even when the true score Y changed. Heise (1969) showed how three measures could yield estimates of reliability and what he called "stability," or the over-time correlation of true scores. Specifically, by assuming that measurement errors at different times are uncorrelated with each other and with the true scores at other time periods, the reliability can be assessed by examining the correlations between waves:

$$\hat{\rho} = \frac{Cor(y_1, y_2)Cor(y_2, y_3)}{Cor(y_1, y_3)} \tag{2}$$

Wiley and Wiley (1970) soon offered an slightly different decomposition, and although a three-wave panel cannot test the assumptions that distinguish Heise's version from that of Wiley and Wiley, both have proven to be very useful in research. Alwin (2007) used both decompositions to estimate reliabilities for all the repeated items in the American National Election Surveys (ANES).

Path analysis is not the only approach to reliability. Here, we propose a multilevel model (Gelman and Hill 2006) that gives a complementary decomposition of the variance in repeated measures. We treat each wave of the three-wave panel as a lower level nested within individuals. For variables with constant true scores, the multilevel model states that the measured variable is the sum of the true score and the measurement error, just as in classic measurement theory:

$$y_{it} = Y_i + \epsilon_{it} \tag{3}$$

for t = 1, 2, 3. We accommodate change over time by adding dummy variables for years:

$$y_{it} = \beta_0 + Y_i + \sum_{t=2008}^{2014} \tau_t + \epsilon_{it} \tag{4}$$

where τ_t equals one for time t and zero otherwise; t = 2006, 2008, 2010, 2012, and 2014. Note that these dummy variables are keyed to actual years. An alternative would index time according to the waves of the panel study. Here, we are more interested in controlling for historical effects, although panel conditioning effects may matter (Halpern-Manners, Warren, and Torche 2016). Each dummy variable adjusts for the change in the mean of y, net of unreliability. The spread of the time dummy variables indicates the accumulated change in that item.

In general, multilevel models decompose total variance in responses into the variance between levels and within levels. In this application, it means variance between persons and within persons. The between-person variance is equivalent to true-score variance ($\hat{\sigma}^2_Y$, net of change captured in the $\hat{\tau}$s), and the within-person variance is equivalent to measurement error variance (σ^2_ϵ). From these we obtain:

$$\hat{\rho} = \frac{\hat{\sigma}^2_Y}{\hat{\sigma}^2_Y + \hat{\sigma}^2_\epsilon} \tag{5}$$

By directly decomposing the residual variance into between- and within-person component parts, we obtain an alternative to Heise (1969).

Many variables in the GSS are ordered but not continuous; many others are nominal. For them we use multilevel ordered logit and multilevel logit models to estimate the variance components (and τs). With ordered and binary logit models, there is no within-person error term in the usual sense. The commonplace identifying restriction sets $\sigma^2_\epsilon = \pi^2/3$; we use that convention and the estimated between-person variance ($\hat{\sigma}^2_Y$) from the estimation to calculate $\hat{\rho} = \hat{\sigma}^2_Y/(\hat{\sigma}^2_Y + \pi^2/3)$.

The GSS Panel Data

In 2006, the GSS drew a panel sample of 2,000 persons from the completed 2006 interviews. Of these randomly selected adults, 1,536 (77 percent) were reinterviewed in 2008; 404 refused and 60 were ineligible (not living in a household or not living in the United States) or deceased. By 2010, 1,276 (64 percent of the the 2,000 original cases) were interviewed for the third time; 211 more refused, 32 more had died, 13 more were not living in a household, and four more were not living in the United States. Similarly, the 2008 cross-sectional sample (*N* = 2,023) was used to create a second panel. Of these 1,581 were reinterviewed in 2010, and 1,295 (again, 64 percent of the original cases) were interviewed a third time in 2012. Of the lost cases, 379 refused the reinterview in 2010, 244 did so in 2012, 61 died, 27 were not living in a household, and 17 left the United States. The same procedures were followed with cases from the 2010 GSS, with similar retention. We pool the data into a combined three-wave, three-panel dataset for this analysis (*N* = 3,875).

Panel members got the same core questionnaire every time. One-time material such as special modules and International Social Survey Programme (ISSP) questions changed from interview to interview, however. The GSS "rotates" some core items (almost all subjective items and a few objective items), putting them on just four of the six ballots in order to maintain time series on more items (see Smith et al. 2015 for details). The number of cases for "rotating core" items, then, is two-thirds of the 3,875 cases minus missing data on that particular item. By way of example, a question about whether a pregnant woman should be able to get a legal abortion "for any reason" (`abany`) is a "rotating core" item; we have 2,344 valid cases for estimating the reliability of that item (1,317 people were never asked that question; 214 were asked but answered "don't know" or gave no reply at least once).

Our focus is on the core items, rotating and otherwise. These are the items that define most of the long-running GSS time series. Using the same wording and the same question order maximizes the opportunity to "measure change by not changing the measure." For a few items, though, changing political context changed the meaning of the question. For example, as we discuss below, one question asked about the respondents degree of confidence in "the people running the executive branch of the federal government" and another about "the people running the Congress." Elections changed the political affiliations of "the people running" those institutions; we use that information to interpret the disappointing reliability results for those two items.

Questions about voting in presidential elections—whether and for whom—usually ask about the most recent presidential election. In a departure from form, panel respondents were asked about the 2004 election in 2010 to obtain a third answer for those voting questions. Respondents in the 2008–2010–2012 panel were asked about the 2008 presidential election in 2012, so their responses were not used in our reliability analysis. Thus, the number of cases used to assess the reliability of the voting and for-whom items was N = 1,246 and 968, respectively.

Choices in the Course of Analysis

Coding, Scales, and Other Measurement Issues

We analyze most variables in the GSS core as they are in cumulative data files (Smith et al. 2015). For continuous variables and ordered variables with more than eight categories, we calculated Pearson correlations in the usual way for the Alwin–Heise estimates (Eq. 2) and fit multilevel regression models to them for the multilevel estimates (Eq. 5). For ordered variables with eight or fewer categories, we calculated polychoric correlations for the Alwin–Heise estimates and used ordered logit regressions for the multilevel model estimates.

Key nominal variables—marital status, employment status, region of origin, religion, and religious origins—were a challenge. Many nominal variables are among the most-used GSS variables, but the nominal distinctions have received far less attention than continuous latent variables in the reliability literature. Even if we think of latent class models, the nature of nominal variables creates inherent contradictions (Alwin 2007:279–280, 287).

We made some dichotomies out of the categories and used tetrachoric correlations for the Alwin–Heise estimates and logit regression for the multilevel model estimates. Table 1 lists the original polytomies and the dichotomies we created. For current religion and religion raised in, we used the classification (which we call `reltrad`) with detailed religious denominations developed by Steensland et al (2000) to construct dummy variables—"TradEv," "TradMain," "TradCath," "TradJew," and "TradOth"—representing conservative Protestants, mainline Protestants, Catholics, Jews, and other religions, respectively. We constructed analogous variables (with 16 appended to the name) for religious origin. `Reltrad` is missing for some cases that are not missing on the standard religion variables (`relig` and `relig16`). For no religion and no religion raised in, we used `relig` to create `none` and `none16`.

We used the most straightforward measurement possible wherever we could. For occupations, we used the new prestige and socioeconomic scores based on the 2012 GSS prestige study (Hauser and Warren 1997; Hout, Smith, and Marsden 2015). We reproduced common scales for support of legal abortion, misanthropy, and gender roles. Rossi's original abortion attitudes scale used six questions;

Table 1: Dichotomies formed from categorical variables.

Categorical variable	GSS mnemonic	Category 1	Category 0	New mnemonic
Race	race	Black	All other	Black
Hispanic origin	hispanic	Hispanic	All other	Hispanic
Marital status	marital	Married	All other	Married
		Never married	All other	Nevermar
Family growing up	family16	Mother & father	All other	Intact16
		Mother only	All other	Singlemom16
Region growing up	reg16	South	Elsewhere	South16
		Outside the U.S.A.	Inside the U.S.A.	Foreign16
Moved since age 16	mobile16	All other	Same city	Moved16
Employment status	wrkstat	Employed	All other	Employed
		Unemployed	Other labor force	Unemployed
		Retired	All other	Retired
Current religion &	relig	No religion	Any religion	None & None16
religion raised in	relig16	Conservative Protestant	All other	TradEv & TradEv16
		Mainline Protestant	All other	TradMain & TradM16
		Catholic	All other	TradCath & TradC16
		Jewish	All other	TradJew & TradJ16
		Other religion	All other	TradOth & TradOth16
Presidential vote	vote04	Voted	Didn't vote	Voted04
	pres04	Bush	Any other	Repvote04
		Kerry	Any other	Demvote04

Note: For details of questions, answer options, and GSS mnemonics, see codebook (Smith et al. 2015).

asking about abortion under any circumstances (abany) is a common extension (e.g., Hout 1999); we estimate the reliability of both scales. Smith (1997) developed a "misanthropy" scale from items about how helpful, fair, and trustworthy people are perceived to be. Finally, four gender-typing items are often used to make a scale (Cotter, Hermsen, and Vanneman 2011). The GSS vocabulary quiz has ten words, though the scale may work better with only seven of them (Malhotra and Krosnick 2007). We assessed the reliability of both the ten-word and seven-word version of the vocabulary score.

We also created some original scales for this study. We combined four questions about suicide to form a suicide scale. We combined Stouffer's (1955) civil liberties items into five scales regarding the freedom of atheists, communists (labeled "red" in figures to save space), militarists, and racists to give speeches, have their books in public libraries, and teach at state universities. Finally, we combine parallel items about socializing with relatives, with friends, with neighbors, and with the patrons of a bar to form a "social life" scale.

Exclusions

We excluded geographical measures that NORC codes from the address the case refers to; these items are part of the administrative record and are not responses to questions the GSS poses to either the respondent or the interviewer. We also excluded aspects of the household context, such as the respondent's relation to the householder and whether the respondent was permanent resident or visitor in the household (we did include several measures of household composition). We excluded information about the interviewer—despite substantive interest in that topic (e.g., An and Winship 2016)—because many cases had different interviewers in different years because of staff turnover, the vagaries of interviewer assignment, and, in a some cases, a move by the respondent.

Ethnic ancestry proved to be too complex to include in this assessment. The GSS asks the question, "From what countries or part of the world did your ancestors come?", records up to three answers, and follows it with the question, "Which one of these countries do you feel closer to?" The complexity of this procedure results in detailed and useful data, but it does not lend itself to either ordered variables or a manageable number of dichotomies. Respondents vary over time in the number of answers they give as well as in their choice of which ancestry they feel closest to. That said, we note that 66 percent of respondents gave the same answer at wave two as wave one, 65 percent gave the same answer at wave three as wave one, 66 percent gave the same answer at wave three as at wave two, and 61 percent gave the same answer in all three waves. Among the 30 percent of respondents who gave the same answer two out of three times, the most common pattern was to give "none" as the different answer. Quantifying this kind of qualitative variation will take a research project of its own.

Classifying Variables by Type and Subtype

Alwin (2007) classified questions as referring to facts, beliefs, attitudes, values, and self-descriptions. The GSS has far more facts and a different mix of attitudes than the surveys he analyzed. We also found the distinctions among beliefs, values, and attitudes hard to apply in the GSS. Instead we created two major types—relatively fixed and not-fixed variables—and subdivided them by topic. We say "relatively fixed" because few things in life are immutable. For most but far from all people, gender and racial identity are relatively fixed. What people say about their parents and upbringing should not change over time, even if the attributes the questions ask about may have changed (e.g., father's or mother's occupation). We expect people to resolve the uncertainty about their upbringing the same way every time. To the extent to which their story of their origins varies, we will detect unreliability.

We subdivided relatively fixed variables into demographic; work; sex, sexuality, and abortion; religious identity and belief; suicide; socioeconomic; trust; civil liberties; health and morale; guns, law, crime, and police; social life; vocabulary; subjective class; politics; public spending; confidence in leadership; recession; gender and family attitudes; and racial attitudes subtypes. These two major types and 19 subtypes are heuristics we use to help us organize and discuss the reliability estimates. The classification we used does not affect the calculations with respect to individual questions; no information about other items of the same type or subtype were used in the estimation. Of course, the averages for the types and subtypes would be different if we reclassified some items.

Results

Summary

Table 2 summarizes our principal results, and Figures 1–8 show important details.[1] The table shows the mean and median reliability under both models for each major type and subtype of item. The figures show the reliability estimates for each item within each subtype. The items are arrayed in most panels of the figures in the order of descending reliability.

The reliability of the core items in the GSS attests to the quality of the data and the survey. Half of the 276 items in the GSS core scored a reliability of 0.74 or higher by the Alwin–Heise method; the median reliability using the multilevel model approach was only slightly lower (0.68). Fixed items were particularly reliable; they averaged just above or below 0.90. The 239 nonfixed items averaged between 0.72 and 0.66.

Relatively Fixed Items

GSS respondents reported the facts of their lives with especially impressive reliability. Relatively fixed aspects of demography and family had reliability of 0.97 or better, as estimated by both the Alwin–Heise method and our multilevel model. Fundamental facts like gender, age, race, Hispanic ethnicity, region of origin, immigrant status, and parents' immigrant status were reported with almost perfect reliability. Age at first birth, the presence of mother and father in the household while growing up, and grandparents' immigrant status were only slightly less reliably reported. People counted their siblings less reliably than other aspects of their basic demography and family origin, but even those items scored close to 0.90.

Religious upbringing was just as reliable as the reporting of demographic facts. People were somewhat less reliable in reporting having been raised in no religion than in each type of religion, but even this estimate was around 0.90.

Questions about parents' educations and occupations varied more; reliability averaged slightly less than 0.80. The occupational pay scores were less reliable than credentials (especially for mothers).[2] The least reliable fixed item is the person's assessment of their parents' income as being above or below average. People reported their parents' work items as reliably as they reported their own. This

Table 2: Summary statistics for reliability estimates by major type and subtype of item.

		Reliability				
		Alwin–Heise		Multilevel		
Major type	*Subtype*	Median	Mean	Median	Mean	Number of items
All		**0.74**	**0.74**	**0.68**	**0.69**	**276**
Fixed (relatively)		**0.94**	**0.90**	**0.92**	**0.89**	**37**
	Demographic	0.99	0.98	0.98	0.98	6
	Religious affiliation	0.99	0.97	0.92	0.94	5
	2004 election	0.99	0.97	0.99	0.97	3
	Family origin	0.97	0.96	0.96	0.93	8
	Socioeconomic origin	0.79	0.81	0.77	0.79	15
Not fixed		**0.71**	**0.72**	**0.65**	**0.66**	**239**
	Demographic	0.98	0.96	0.93	0.91	8
	Work	0.91	0.89	0.77	0.79	17
	Sex, sexuality, & abortion	0.86	0.82	0.83	0.78	18
	Religious identity & beliefs	0.85	0.85	0.81	0.81	19
	Suicide attitudes	0.83	0.82	0.74	0.75	6
	Socioeconomic status	0.79	0.82	0.79	0.79	14
	Trust & misanthropy	0.76	0.74	0.68	0.68	4
	Civil liberties	0.74	0.73	0.63	0.65	20
	Health & morale	0.73	0.71	0.64	0.62	7
	Guns, laws, crime, & police	0.72	0.76	0.68	0.71	14
	Social life	0.70	0.71	0.61	0.64	8
	Vocabulary	0.68	0.67	0.68	0.63	12
	Subjective class & mobility	0.68	0.63	0.62	0.59	5
	Politics & government	0.67	0.68	0.62	0.63	10
	Public spending	0.64	0.64	0.53	0.54	17
	Confidence in leadership	0.62	0.61	0.53	0.52	14
	Recession experience	0.59	0.60	0.49	0.48	8
	Gender & family attitudes	0.59	0.57	0.52	0.52	17
	Race & immigration attitudes	0.51	0.53	0.47	0.47	21

Note: For details on questions, answer options, and GSS mnemonics, see codebook (Smith et al. 2015).

generalizes the finding in previous studies that men's reports of their father's attributes are almost as reliable as their reports of their own (Bielby, Hauser, and Featherman 1977a, 1977b; Alwin 2007).

People very consistently reported their voting behavior in the 2004 election. All three reliabilities were well above 0.90. All the other kinds of items involved recalling events and statuses over longer time than the election items did.

The Alwin–Heise approach distinguishes between reliability and stability, the latter simply being the correlation between consecutive measurements. The stability of relatively fixed items should be very close to 1.0. Major departures from unity call the notion of "fixed" into question. The 37 relatively fixed items generated 74 stability estimates (two per item; one for between waves one and two and

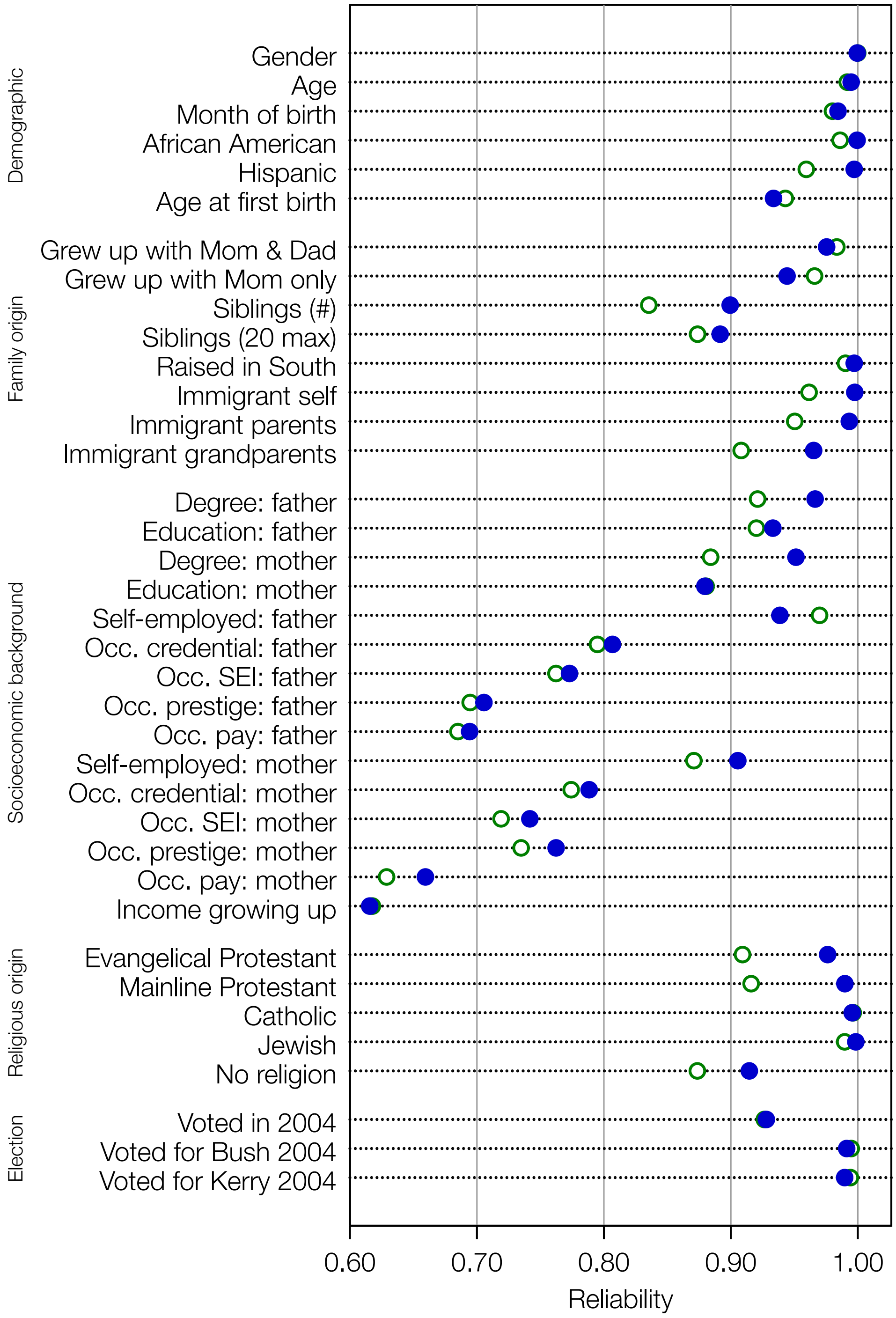

Figure 1: Reliability estimates for relatively fixed items: demographic, family, socioeconomic, religious, and election items.

Key: The estimates from the Alwin–Heise model are shown by solid blue circles; the estimates from our hierarchical linear models are in open green circles. The open green circle is not visible when the estimates are very close.
Notes: The estimates pool data from all three three-wave panels, except for the election items, which are from the first three-wave panel (2006–2010) only.
Source: Authors' calculations from the General Social Survey three-wave panels, 2006–2014.

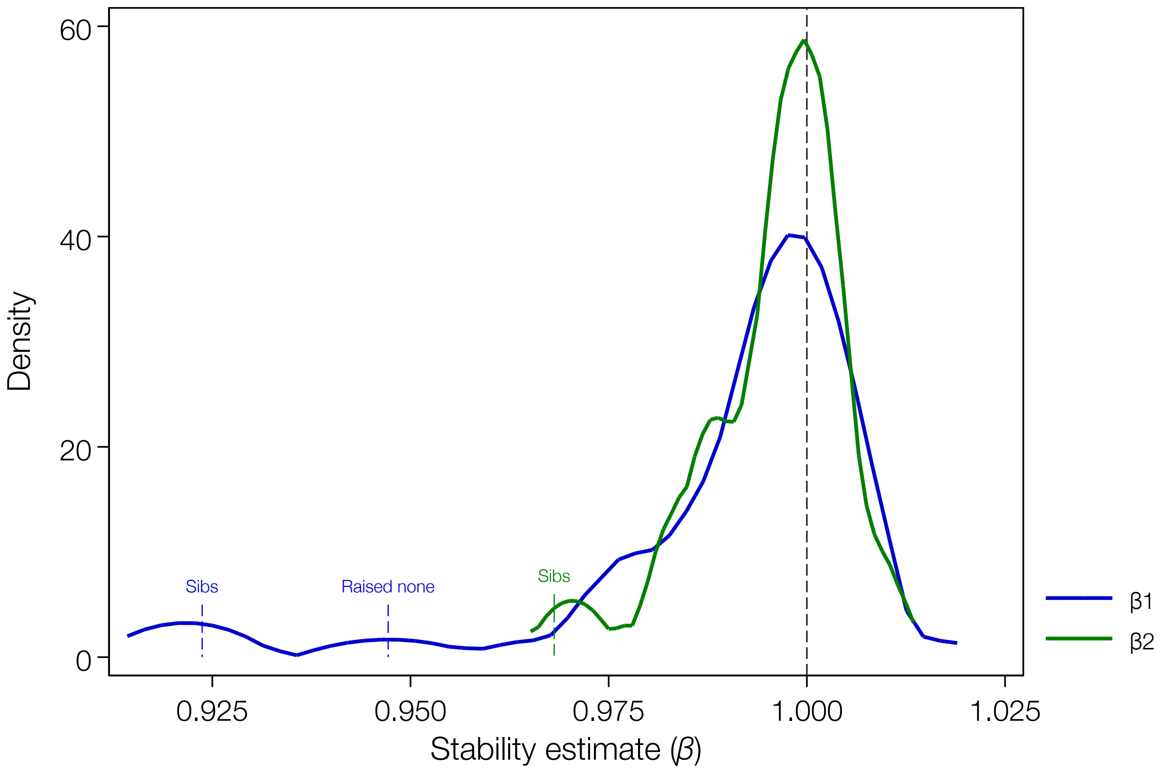

Figure 2: Kernel density curves for stability estimates between the first and second waves (β_1) and second and third waves (β_2) for relatively fixed items.

Notes: The three items with βs less than 0.975 are labeled. The vertical line at 1.0 is the theoretically derived value that a stability item would have if the item was, truly, fixed.
Source: Authors' calculations from the General Social Survey three-wave panels, 2006–2014.

one between waves two and three). The estimates clustered very close to the hypothesized value of 1.0, as shown in kernel density plots in Figure 2. All but three were in the range from 0.980 to 1.020. The three lowest stability estimates were for siblings (β_1 = 0.924), being raised with no religion (β_1 = 0.948), and siblings again (β_2 = 0.971). These estimates confirm that the items that appear relatively fixed on their face were, in fact, nearly perfectly fixed, except for the small unreliability reported in Figure 1. The dummy variables from the multilevel models confirmed this high degree of stability; none of the 37 sets of coefficients were statistically significant.

Not-Fixed Demographic, Socioeconomic, and Work Items

We begin our analysis of items that might change over time with demographic and socioeconomic facts (see Figure 3).[3] Demographic and socioeconomic fundamentals including marital status, children ever born, geographical mobility, home ownership, own degrees, and spouse's degrees were perfect or nearly so (reliability very close to 1.0). Children ever born, geographic mobility, years of education, and spouse's years of education were also highly reliable; estimates exceeded 0.90. The least reliable among this set of items was the respondent's guess about how segregated the neighborhood was (Alwin–Heise reliability of 0.85 and multilevel estimate of 0.75).

The reliability of more complicated socioeconomic variables was slightly lower. Occupational status is based on verbatim descriptions of job title and principal duties, which are then assigned a census occupational code. A score is then assigned, depending on the code. We used four scores: the degree associated with the occupation, its pay, the weighted average of the two (socioeconomic index [SEI]), and its prestige as assessed in popular ratings (Hout et al. 2015). These multistep assessments (that began with people's descriptions in their own words) resulted in reliability estimates clustered around 0.75. Income was reported in categories (see Smith et al. 2015). Taking the categories as an ordinal scale or transforming them to midpoints and taking logarithms (to create a ratio scale) yields reliability estimates between 0.75 and 0.80.

Subjective socioeconomic variables were harder for people to answer reliably. Subjective social class was the most reliable, with a reliability of 0.71. Intergenerational mobility items asking people to compare themselves with their parents and children rated 0.67 and 0.70 using the Alwin–Heise method and lower using the multilevel model. People's sense of their relative income was almost exactly as reliable as their sense of their parents' income (0.61). The item asking people if hard work or luck matters more in advancement scored the worst among socioeconomic variables (0.65).

Most work items were highly reliable, rating 0.85 or higher. People reported their employment status, union membership, government employment, and self-employment very reliably. They were slightly less reliable reporting weeks worked and whether they supervised others or were supervised. Even worse were whether their supervisors or subordinates had supervisors and subordinates, the size of their workplace, and how segregated it was.

Worst among the work items were those most affected by the Great Recession of 2007–2009 and the slow recovery that followed it. The layoffs and disruptions associated with the recession should, theoretically, be reflected in the stability estimates for these items. But they also seem to have affected the reliability estimates for being unemployed or out of work,[4] hours worked, and subjective assessments of family finances and job prospects. This is troubling because these statuses obviously did change—and in the case of unemployment and, more generally, out of work, changed back—through the course of the recession and slow recovery. We conjecture that the estimates would be substantially higher—on par with employment status—in a panel during a more stable economic time.

Not-Fixed Religious Identity and Beliefs and Suicide Items

People very reliably reported their current religion and religious behaviors such as praying and attending religious services (see Figure 4). Current religion and religious origin were generally very close in reliability (compare with Figure 1).

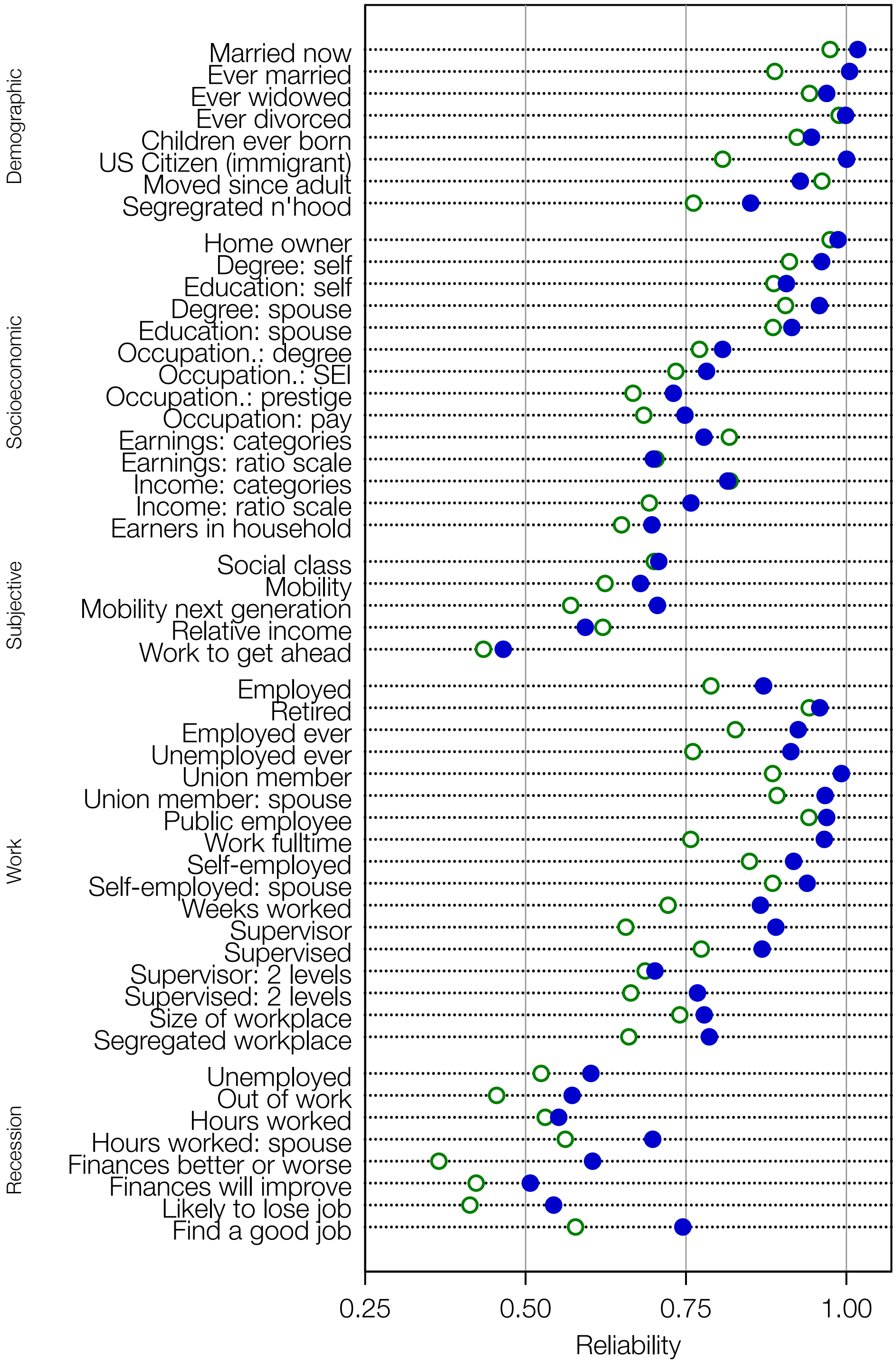

Figure 3: Reliability estimates for items about demographic, socioeconomic, and work facts plus subjective socioeconomic items.

Key: The estimates from the Alwin–Heise model are shown by solid blue circles; the estimates from our hierarchical linear models are in open green circles. The open green circle is not visible when the estimates are very close.
Notes: The "Recession" items are mostly from the "Work" category. We collected them into a special subtype to highlight the effects of the recession of 2007–2010.
Source: Authors' calculations from the General Social Survey three-wave panels, 2006–2014.

People were less reliable in reporting their religious activities other than attending services (`relactive`) than in reporting prayer and attending services. They were also less reliable in reporting the strength of their religious identity than of their denominational group. We suspect that much of that unreliability came from the unprompted but consistently coded "somewhat strong" response. This intermediate code was mainly used once in the three interviews by respondents who ever used it. Reliability of this item would likely improve if "somewhat strong" was read as part of the question. Changing the item in this way could, of course, affect the time series of strength of religious identification.

The relatively high reliabilities of voting (Figure 1) and church attendance (Figure 4) seem somewhat inconsistent with evidence of bias in survey reports of these behaviors. Survey estimates of voter turnout generally exceed official turnout, and Hadaway, Marler, and Chaves (1993) estimated that half the Protestants who reported attending religious services on a given weekend were not present at any local churches.[5] Reliability does not rule out bias in these or other items. Reliability is, in some sense, a case of "I told my story, and I'm sticking to it." That is, people may be reliable in both their behavior and their account of it, even though there is a discrepancy between the two. That interpretation rests on the assumption that people remember what they said two years earlier. We think it is pretty unlikely that such an assumption could be true of most respondents. Instead, it might be that surveys select for voters, as Silver and Anderson (1986) note, and church attenders. The people who skip voting and church services may also exempt themselves from surveys. If so, then the GSS (and other surveys) could be right about the respondents but wrong about the population. With a response rate of slightly more than 70 percent, the GSS covers the population better than most surveys, but if nonvoters and people who seldom or never attend religious services are among the thirty percent it misses, then the sample, not the items, may be part of the bias.

One hint that this interpretation is at least partially correct can be seen in the differences among those who participated in all three waves of the GSS panel, those who did two interviews but not the third, and those who participated in just the first interview. In the 2006–2008–2010 GSS panel, 68 percent of people who completed all three interviews reported having voted in the 2004 presidential election when first interviewed in 2006. Of those who did two but not all three interviews, 62 percent reported having voted in the 2004 presidential election; of those who did the first interview but neither reinterview, 58 percent reported voting in the 2004 presidential election. Extrapolating from that pattern to the people who refused even the first interview, we can guess that barely half would have reported voting in 2004.

Religious beliefs were among the most reliable subjective items in the GSS. Core beliefs about life after death and the existence of God were particularly reliable. Views of the Bible and identity as a religious or spiritual person were somewhat less reliable. Catholics' views of the infallibility of the Pope (on doctrinal matters) was the least reliable religious belief. Some of this lower reliability might be attributable to the change from Pope Benedict XVI to Pope Francis in 2013 (which changed the underlying referent in question), but only respondents in the 2010–2012–2014 GSS panel experienced this change. In an analysis of only the 2006–2008–2010 panel,

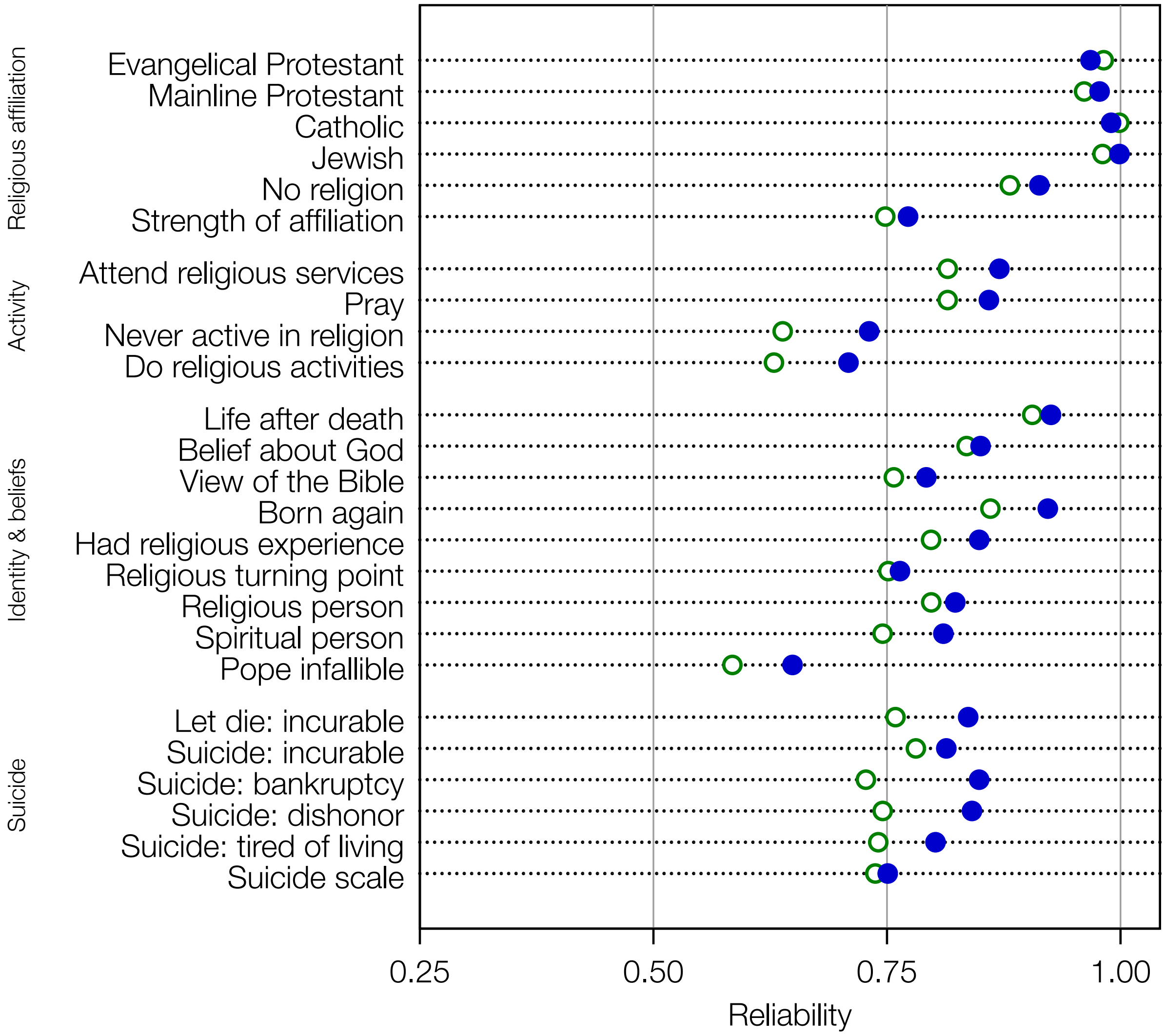

Figure 4: Reliability estimates for items about religious facts, identities, and beliefs plus items about suicide and end-of-life.

Key: The estimates from the Alwin–Heise model are shown by solid blue circles; the estimates from our hierarchical linear models are in open green circles. The open green circle is not visible when the estimates are very close.
Source: Authors' calculations from the General Social Survey three-wave panels, 2006–2014.

which experienced no change, views on the infallibility of the pope still had the lowest reliability among religious beliefs (Hout and Hastings 2012).

Attitudes toward the end of life and suicide were very reliable, especially compared to the reliability of items on other controversial topics. Interestingly, the scale formed by summing the four suicide items was not as reliable (by the Alwin–Heise method) as the separate items.

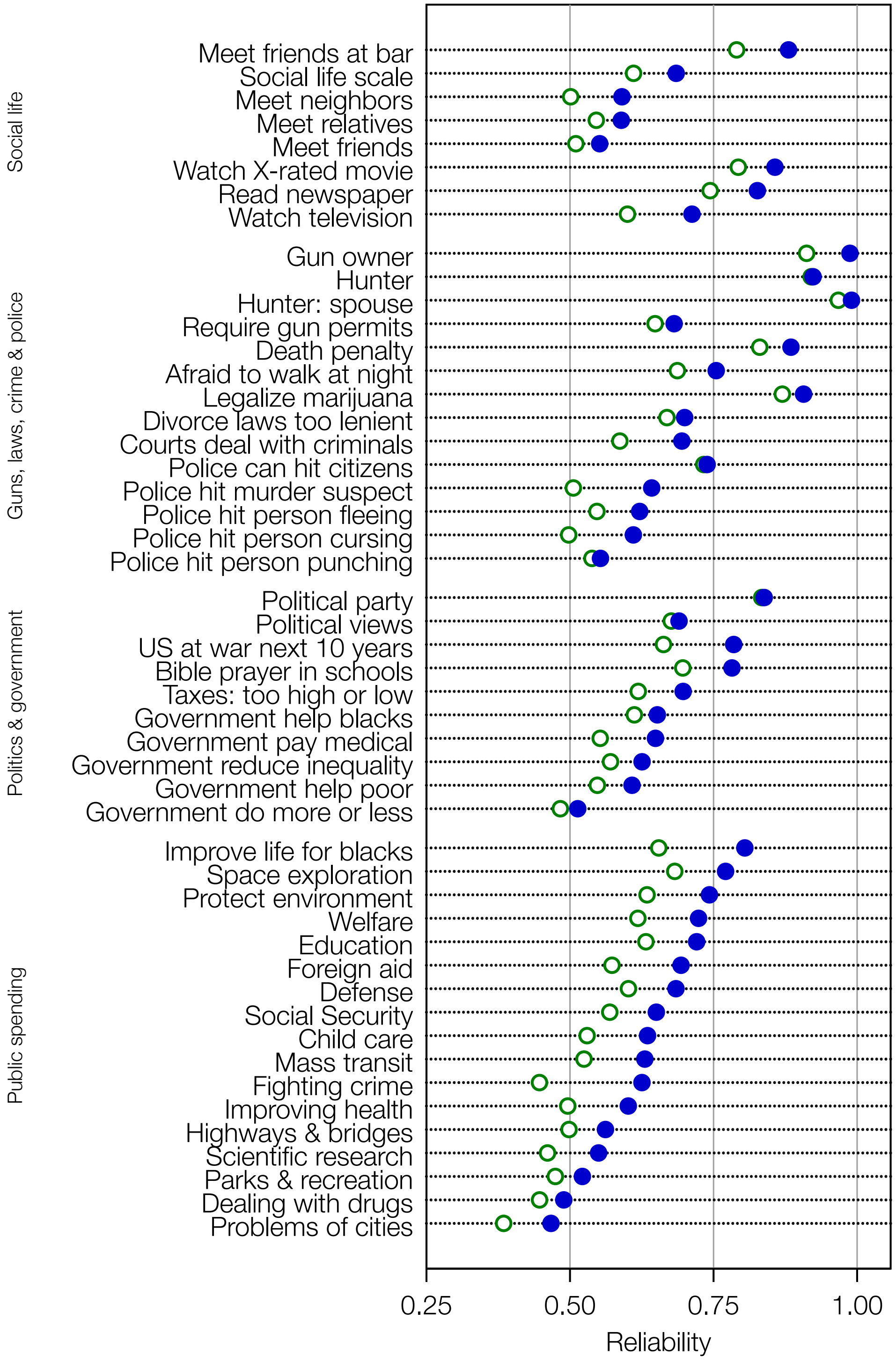

Figure 5: Reliability estimates for items about social life, politics, and government.

Key: The estimates from the Alwin–Heise model are shown by solid blue circles; the estimates from our hierarchical linear models are in open green circles. The open green circle is not visible when the estimates are very close.
Source: Authors' calculations from the General Social Survey three-wave panels, 2006–2014.

Not-Fixed Social, Political, and Government Items

Most activity reports in the GSS rated high on reliability. The social life items were exceptions (see Figure 5). Meeting friends in a bar had high reliability; the rest—spending social evenings with relatives, neighbors, and coworkers—scored much lower (around 0.60) than most activities did. The four-item scale was more reliable, scoring 0.70 by the Alwin–Heise method and 0.65 by the multilevel model. The media use items were much more reliable, especially the X-rated film and newspaper items.

People reported their gun ownership and hunting as reliably as they reported basic demographics (that is, almost perfectly). Spouse reports were almost as reliable as self-reports.

Attitudes toward capital punishment and legalizing marijuana were among the most reliable controversial opinions. Requiring gun permits, fear of crime in the neighborhood, divorce laws, and lenient courts were less reliable than capital punishment and marijuana but still relatively reliable, with Alwin–Heise estimates greater than 0.65. Views about the police hitting citizens, even those in custody, were the least reliable in this subject area, perhaps because of deaths of African American men at the hands of the police receiving more publicity starting in 2010 and of the Black Lives Matter movement beginning in 2013, changing opinions.

Political party identification is a composite variable constructed from a basic question and a follow-up that varies according to the person's answer to the basic question. The composite (`partyid`) was as reliable as most facts (0.84). Political views from liberal to conservative had substantially less reliability (0.66) than party. Political items about taxes and government action on inequality and health were slightly less reliable than political views. The composite item about whether government should be more or less active in solving the country's problems (`helpnot`) was the least reliable among these political items.

The reliability of attitudes toward public spending were spread across a broad range from 0.4 to 0.8. Whether asked in the main or alternative form (containing longer explanations of issues; not shown in Figure 5) had much less impact on reliability than the subject itself did. Answers to questions regarding spending on racial disparity, the space program, the environment, education, welfare, foreign aid, and defense were relatively more reliable than attitudes toward spending on science, crime, drugs, and urban issues. The alternative form of these less reliable items yielded more reliable answers, suggesting that the difference between more reliable and less reliable spending items is information or familiarity.

Not-Fixed Civil Liberties, Trust, and Confidence Items

The famous Stouffer civil liberties items have an average reliability of 0.70. The six questions Stouffer (1955) included in his original analysis—the atheist and communist items—were among the more reliable items, as were the ones about "an admitted homosexual." Items about speech were somewhat more reliable than the ones about library books and college professors, except when the subject was someone advocating military overthrow of the government (shown in Figure 6).

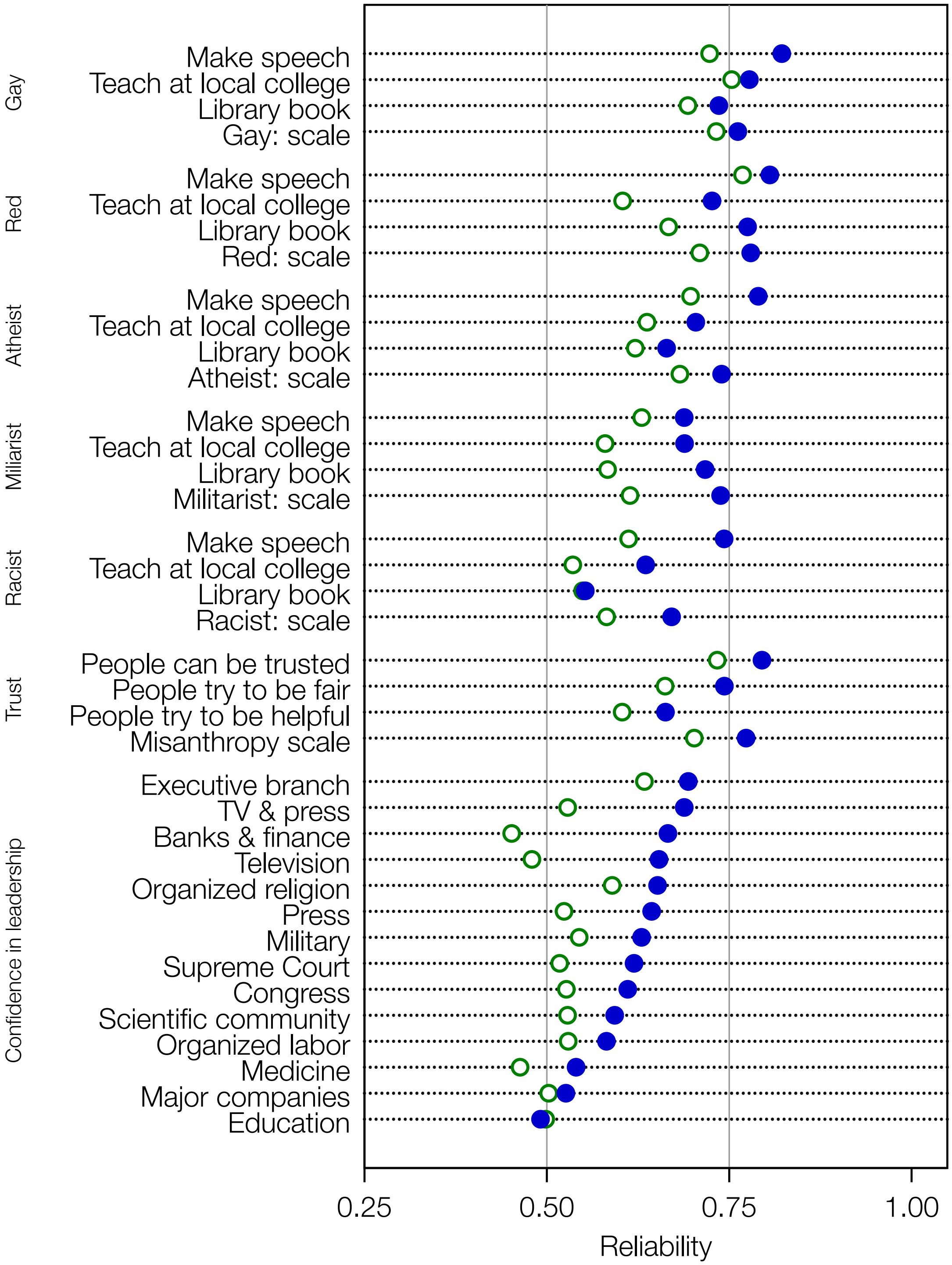

Figure 6: Reliability estimates for items about civil liberties, trust, and confidence in institutions.

Key: The estimates from the Alwin–Heise model are shown by solid blue circles; the estimates from our hierarchical linear models are in open green circles. The open green circle is not visible when the estimates are very close.
Notes: The data for the executive branch was limited to the 2010–2014 panel because the executive branch changed from Republican to Democrat during the other two panels.
Source: Authors' calculations from the General Social Survey three-wave panels, 2006–2014.

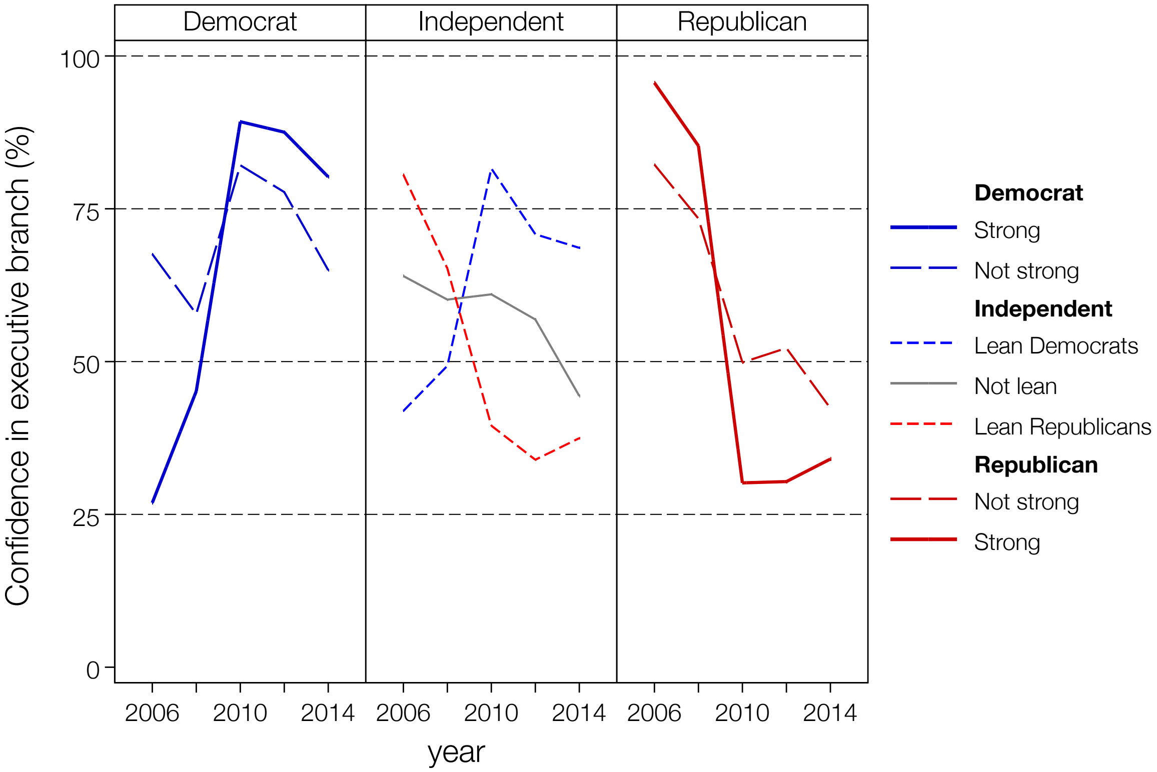

Figure 7: Confidence in people running the federal government by year, political party preference, and strength or leaning of party preference.

Key: The estimates from the Alwin–Heise model are shown by solid blue circles; the estimates from our hierarchical linear models are in open green circles. The open green circle is not visible when the estimates are very close.
Notes: Means are adjusted for within-person stability and variability.
Source: Authors' calculations from the General Social Survey three-wave panels, 2006–2014.

The misanthropy scale (Smith 1997) and its constituent items were fairly reliable. Answers about confidence in social and political institutions varied in their reliability estimates near 0.7 down to 0.5. Answers about the executive branch were a special case because the referent changed when Obama succeeded Bush in January of 2009. Answers flip-flopped dramatically at the change of administration as shown in Figure 7; Democrats predictably felt more confident and Republicans just as predictably less confident in the people running the executive branch after the Democrats moved in and the Republicans moved out. The flip was greatest among strong partisans, less among weak partisans, and almost as great among independents with partisan leanings as among weak partisans. Consequently, the reliability estimates from the 2006–2008–2010 and 2008–2010–2012 panels were not usable.[6] The estimate of 0.68 shown in Figure 6 was the estimate from the 2010–2012–2014 panel (the one with no change of administration). Confidence in people running major companies (CEOs) and education were least reliably reported. The recession may have affected confidence in CEOs, but it is hard to see why it might have affected the reliability of ratings for leaders in education. We got similarly low reliability estimates for these two institutions in all three panels when we separated them, making it seem less likely that the low reliability reflected disruptive events. We conclude that the questions about confidence in people running major companies and education were just less reliable than the others.

Not-Fixed Items About Health, Morale, Sex, Gender, Family, Race, and Vocabulary

Self-rated health and marital happiness were very reliable for subjective items (0.76). General happiness and other satisfaction items were significantly less reliable. Like the employment status items, satisfaction with one's finances and employment might have been destabilized by the recession; more analysis of the flux in these measures is part of our agenda for the future.

Questions about abortion, sex, and sexuality yielded very reliable answers despite the controversies surrounding the topics (or perhaps because of the polarizing effects of those controversies). With reliabilities mostly above 0.80, these hot-button issues represent rather "mature" attitudes that were reliably reported (Hout 1999). Other controversial aspects of sexual politics were also quite reliable. Identity as lesbian, gay, bisexual (LGB), or exclusively heterosexual was very reliable, as were attitudes toward same-sex sex and premarital sex. Attitudes toward sex involving adolescent partners and extramarital sex were less reliably reported. The items about restricting pornography and birth control for adolescents were the least reliable among this battery of items.

Attitudes toward gender roles and other family issues, some of them part of the GSS from beginning, were less reliable than the more political items. The questions that comprise the gender-roles scale (Cotter et al. 2011) were pretty good, scoring around 0.65, as did the four-item scale itself. The question asking people to rank five attributes that might be desirable for a child to have did poorly on the whole, mainly because people unreliably placed hard work relative to the others. They most reliably placed "to obey."

Racial attitudes were far less reliable than other beliefs and attitudes. The 11 best items were moderately reliable ($\approx$0.50); crucial measures of social distance and racial stereotypes were quite unreliable (ranging from 0.25 to 0.40). The low reliability of these items will be a matter of great concern for scholars who need to measure these aspects of racial attitudes. Perhaps the historic 2008 election cycle shifted public thinking. But the unreliability of these measures was not accompanied by net change, just flux within persons from wave to wave of the panel. The stereotype items are actually more highly correlated within waves than between; for example, the items on how intelligent blacks and whites are in the same wave correlate more highly (r = 0.40) than the respondent's rating of black intelligence in one interview and the next (r = 0.22).

Most of the unreliability of the racial stereotypes items seems to be in the details. These items asked respondents to rank blacks and whites on seven-point scales. Between 49 and 55 percent of respondents scored the groups in the middle category (a score of 4). Variation around that appears to have been much less reliable than the

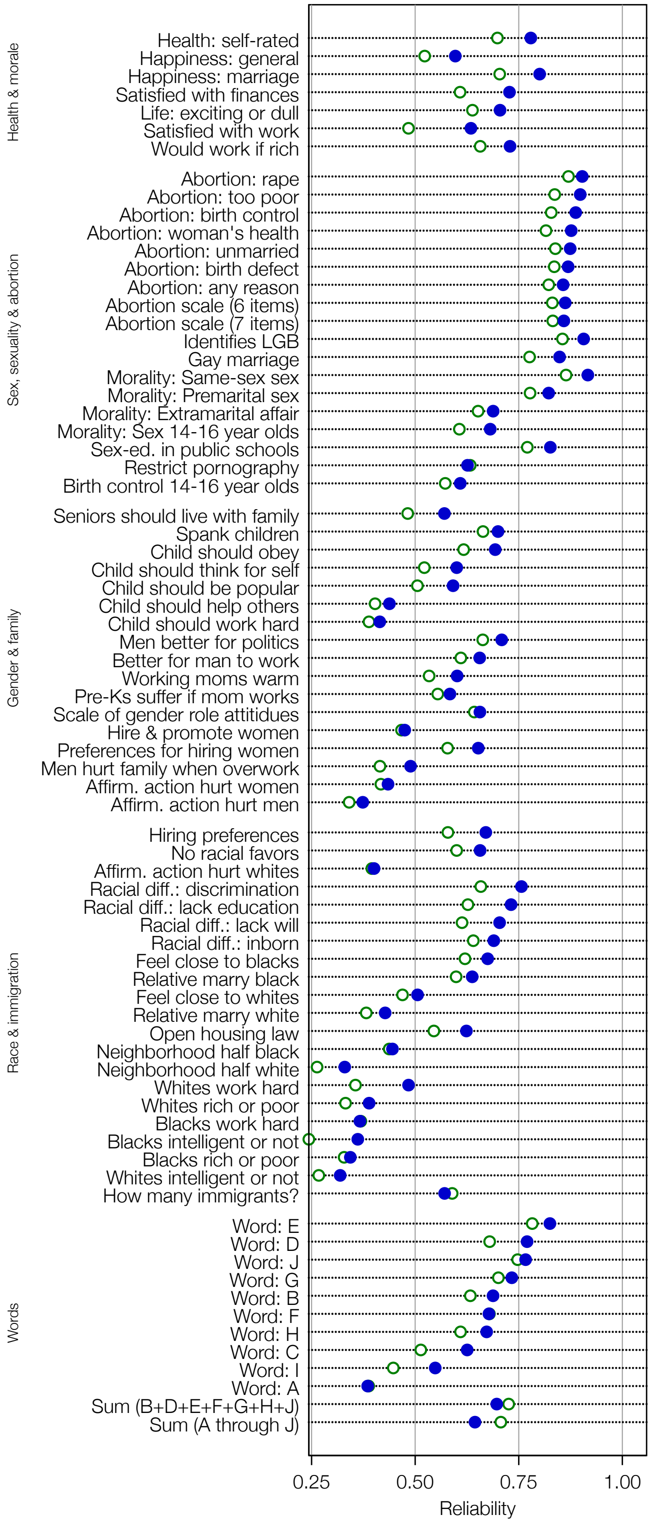

Figure 8: Reliability estimates for items about sex, gender, race, and vocabulary.

Key: The estimates from the Alwin–Heise model are shown by solid blue circles; the estimates from our hierarchical linear models are in open green circles. The open green circle is not visible when the estimates are very close.
Source: Authors' calculations from the General Social Survey three-wave panels, 2006–2014.

decision to depart on the high side or low side of the middle when the item referred to blacks. Treating scores 1–3 as low, 4 as medium, and 5–7 as high and ignoring variation within high or low substantially increased the estimated reliability. For example, the reliability of the three-point "intelligent" item was 0.65 for blacks (but it barely raised reliability for whites, just 0.36). Reducing the items to dichotomies 1–4 versus 5–7 further increased the reliability of the black "intelligent" stereotype to 0.78 and the white "intelligent" stereotype to just 0.61.

The 10-item vocabulary quiz had an acceptable reliability of 0.65. Some of the individual words were even more reliable than the quiz score itself, but three (words C, I, and especially A) had reliability.[7] The variation is not related to the difficulty of the words. The two hardest words—C and J—had reliabilities of 0.64 and 0.75, respectively; the two easiest words—A and F—had reliabilities of 0.39 and 0.73, respectively. Malhotra and Krosnick (2007) addressed the reliability of the words in the GSS vocabulary quiz. Because several words have about the same degree of difficulty—as gauged by the proportion correct—vocabulary might be measured more reliably by picking a subset of words that vary more in difficulty. Their alternative scale A (words a, c, h, and i) had a reliability of 0.53 in the GSS panel; their alternative scale B (words d, e, f, and g) had a reliability of 0.72. The first is substantially less reliable than the standard ten-item scale, but scale B offered a modest improvement. A third alternative composed by dropping the three least reliable words could also raise reliability to 0.72.

In summary, the GSS core items were impressively reliable on the whole, but there was substantial variation in the degree of reliability. People report the facts of their lives very reliably. Important beliefs (about God, for example) and identities (as sexual orientation and political party) are almost as reliable as facts. Attitudes on topical subjects (abortion, gay marriage, and marijuana) are also impressively reliable. Complicated, subtle, unfamiliar, and volatile issues are substantially less reliably reported. The least reliable items were reports of racial stereotypes. We found that reliability for them could be increased by simplifying them; apparently, variance below and above the mid-point is mostly random.

Change and Stability

The main interest in panel data, is, of course, measuring social change. Our approach here is to focus on what might be considered "structural change"—shifts over time in the means of continuous variables and in the marginal distributions of categorical variables. The historical events that occurred from 2006 to 2014, especially the Great Recession (and its aftermath) and the election (and reelection) of President Obama, make structural change both likely and important. Full engagement with these major events and their consequences would merit research articles of their own. We expect the research community to take up these topics. Substantive articles published to date include articles on health care (Brooks and Manza 2013), trust and confidence (Owens and Cook 2013), redistribution of wealth and incomes (Owens and Pedulla 2014), and religious identification (Hout and Fischer 2014).

Table 3: Estimates of change parameters (τs) from multilevel models for selected GSS variables: 2006–2014.

		Change parameters (τs)						
Description	Model	2008	2010	2012	2014	σ_u^2	ρ	N
Age: 2006-08-10	r	1.98	3.99	—	—	268.86	0.99	1,274
		(0.05)	(0.05)			(10.67)		
Age: 2008-10-12	r	—	1.94	3.97	—	269.72	0.99	1,294
			(0.08)	(0.08)		(10.67)		
Age: 2010-12-14	r	—	—	1.95	3.93	262.40	0.99	1,304
				(0.04)	(0.04)	(10.67)		
Unemployment	l	0.10	1.13	0.79	0.59	3.85	0.54	2,956
		(0.24)	(0.22)	(0.24)	(0.27)	(0.61)		
Weeks worked last year	r	0.41	−2.21	−2.70	−2.66	383.60	0.72	3,875
		(0.44)	(0.44)	(0.51)	(0.59)	(147.16)		
Hours worked	r	−0.53	−1.68	−2.30	−1.46	110.19	0.51	2,818
		(0.50)	(0.49)	(0.55)	(0.64)	(4.44)		
Financial prospects	o	−0.66	−0.75	−0.84	−0.73	2.41	0.42	2,602
		(0.09)	(0.08)	(0.09)	(0.11)	(0.14)		
Confidence in banks	o	−0.74	−2.10	−2.04	−1.47	2.71	0.45	2,601
		(0.10)	(0.10)	(0.11)	(0.12)	(0.17)		
Political party identification	o	−0.38	−0.23	−0.18	−0.21	23.42	0.88	3,861
		(0.08)	(0.08)	(0.09)	(0.10)	(0.95)		
Spending on health	o	−0.30	1.13	0.87	1.12	3.25	0.50	1,899
		(0.14)	(0.14)	(0.14)	(0.16)	(0.27)		
Spending on environment	o	0.21	1.07	1.10	0.96	5.78	0.64	1,897
		(0.14)	(0.14)	(0.15)	(0.17)	(0.43)		
Legalize marijuana?	l	0.21	0.88	1.21	2.01	23.31	0.88	2,588
		(0.16)	(0.16)	(0.18)	(0.21)	(2.73)		
Gay marriage	o	0.21	0.59	0.96	1.54	11.56	0.78	2,558
		(0.09)	(0.09)	(0.11)	(0.13)	(0.58)		

Notes: Change measured relative to 2006. Standard errors in parentheses. Models key: regression (r), logit regression (l), and ordered logit regression (o).

Source: Authors' calculations from the General Social Survey three-wave panels, 2006–2014.

To validate our approach to reliability in the midst of change, we start with an example in which the outcome is known in advance, just to verify that time dummy variables capture the changes all persons share. Chronological aging is as predictable as a social variable can get. Each person interviewed as part of the GSS panel was two years older when reinterviewed and four years older at the time of the third and last interview. If our model works, then with age as the outcome variable, the coefficients for the time dummies will increase by 2 each year. As the first three results in Table 3 show, the multilevel model performs as expected; the change parameters, rounded to whole numbers, are estimated to be 2 and 4 for waves two and three from each panel.

Having validated the approach, we turning to substance. We consider four variables likely to have been strongly affected by the Great Recession. Unemployment was the personal experience that most nearly defined the Great Recession (Grusky, Western, and Wimer 2011). The GSS measures employment status with the same question as the Current Population Surveys; the Bureau of Labor Statistics uses the answers to that question to calculate the official unemployment rate. In the GSS, unemployment rates for panel cases were 5.1 percent, 5.1 percent, 9.7 percent, 7.7 percent, and 6.7 percent in 2006, 2008, 2010, 2012, and 2014, respectively.[8] Including just people in the labor force according to the practice at the Bureau of Labor Statistics, we regressed unemployment status on year using the multilevel logit regression model. The odds of being unemployed were not significantly higher yet in 2008, nearly tripled in 2010 (exp(1.13) = 3.10), were still more than double (exp(0.79) = 2.20) in 2012, and still 81 percent higher in 2014 (exp(0.59) = 1.81). These changes in overall levels were impressive and very much in line with the trends in the national economy.

Yet employment status appears to one of the least reliable facts in the GSS, whether we look at the multilevel model estimate or the Alwin–Heise estimate. People laid off in the Great Recession had an even more extreme outcome than expected based on the time dummies in the model. The absolute change from employed to unemployed exceeded expectations. Most went back to being employed after their unemployment spell. That change also exceeded the modest expectations based on that year's time dummy. Both models mistake accurately reported volatility for unreliability. Our dummy variables are insufficient to the task of fully adjusting for structural change.

Weeks and hours worked declined among those who were employed during the recession. The multilevel model estimates that Americans worked 2.2 fewer weeks in 2009 than 2005, and they had not recovered those weeks of work by 2013.[9] Similarly, the GSS panel recorded 1.5 to 2.3 fewer hours of work in the last three waves than in the first two. But as with unemployment, substantial within-person volatility comes through here as unreliability on top of the aggregate year-to-year changes. Weeks worked appears to be substantially more reliable than either unemployment status or hours worked.

Subjective feelings about the economy were affected as well (Hout and Hastings 2014). The `goodlife` question asks people if they agree or disagree that "The way things are in America, people like me and my family have a good chance of improving our standard of living." The odds of offering a positive answer fell sharply between 2006 and 2008, fell a little more in 2010, and even more in 2012; the $\hat{\tau}$s are –0.66, –0.75, and –0.84, respectively, before rebounding ever so slightly in 2014. The GDP recovery in 2010 failed to register either in employment or people's assessments of their prospects.

The banks and financial institutions were implicated early in the liquidity crisis that precipitated the Great Recession. Americans' confidence in banks and financial institutions declined significantly already in 2008, and it continued downward through 2012. In 2014, there was a significant increase in confidence in the people running the banks (as the question phrases it), but the rebound is only one-third of the way back to its prerecession level. As with the other recession-related items, within-person variance not captured by the year effects comes through as unreliability.

Politics were also in flux throughout the 2006–2014 period. The populace swung strongly toward the Democrats between 2006 and 2008 ($\hat{\tau}$ = –0.38), but the Democrats lost significant ground between 2008 and 2010 ($\hat{\tau}$ = –0.23). Party identification did not change significantly after 2010. Americans still leaned significantly more to the Democrats in 2014 than in 2006, but the net shift was much smaller than the immediate change between 2006 and 2008. Nor was there nearly as much change in underlying political views. A parallel analysis of where people placed themselves on the liberal–conservative spectrum (not shown) indicated that Americans' views moved to the left briefly (the year effect for 2008 was significant) but moved back to status quo ante and has remained there. These results are in accord with the findings of political scientists that party identification is "remarkable" because, though it is a less central identity than gender, race, or hometown, it is just as stable within individuals (Green, Palmquist, and Schickler 2002:2–6).

Most Americans (80 percent) felt that the government was not spending enough to improve the nation's health in 2006. Passage of the Affordable Care Act (Obamacare) and other developments dramatically reduced that percentage of Americans who think the government spends "too little" on "improving the nation's health" from 80 percent to 60 percent between 2008 and 2010. Few go so far as to say the United States spends "too much" on the nation's health. Peoples' views of most kinds of spending tempered during this time. Table 3 includes spending on the environment as an example. This and several other spending items moved from "too little" to "right amount" during the Obama administration.

While spending views moved to the center, Americans of all political persuasions dramatically increased their support for legalized marijuana and gay marriage between 2006 and 2014. The $\hat{\tau}$s were 0.21, 0.88, 1.21, and 2.01, respectively, over time from 2008 to 2014. In percentage terms, approval rose from 37 percent in 2006 to 60 percent in 2014. Support for a legal right for gay couples to marry grew almost as fast. The $\hat{\tau}$s were 0.21, 0.59, 0.96, and 1.54, respectively. In percentage terms, agreement rose from 35 percent in 2006 to 38 percent in 2008 to 59 percent in 2014. On both issues, majorities supported the liberal position in both 2012 and 2014.

And that is how the answers to some of the apparently less reliable GSS questions by psychometric reckoning turn out to be, nonetheless: useful and at least somewhat predictable.

Most items were much less volatile over time. To quantify that statement, turn to the βs from the Alwin–Heise model as estimated for items that are not fixed. As in Figure 2, we calculate kernel densities for both βs; they are graphed in Figure 9. The βs center somewhat below 1.0 (more so for β_1) and skew left (more below 1.0 than above it). The recession items were notably low by this reckoning, as were some items related to race, crime, and confidence in institutions. Figure 10 displays both βs for items that had at least one β less than 0.8. Recall that the stereotype items stood out as having very low reliability. They also turned out to have very low stability as well. The item about black intelligence, in particular, was very unstable by these estimates.

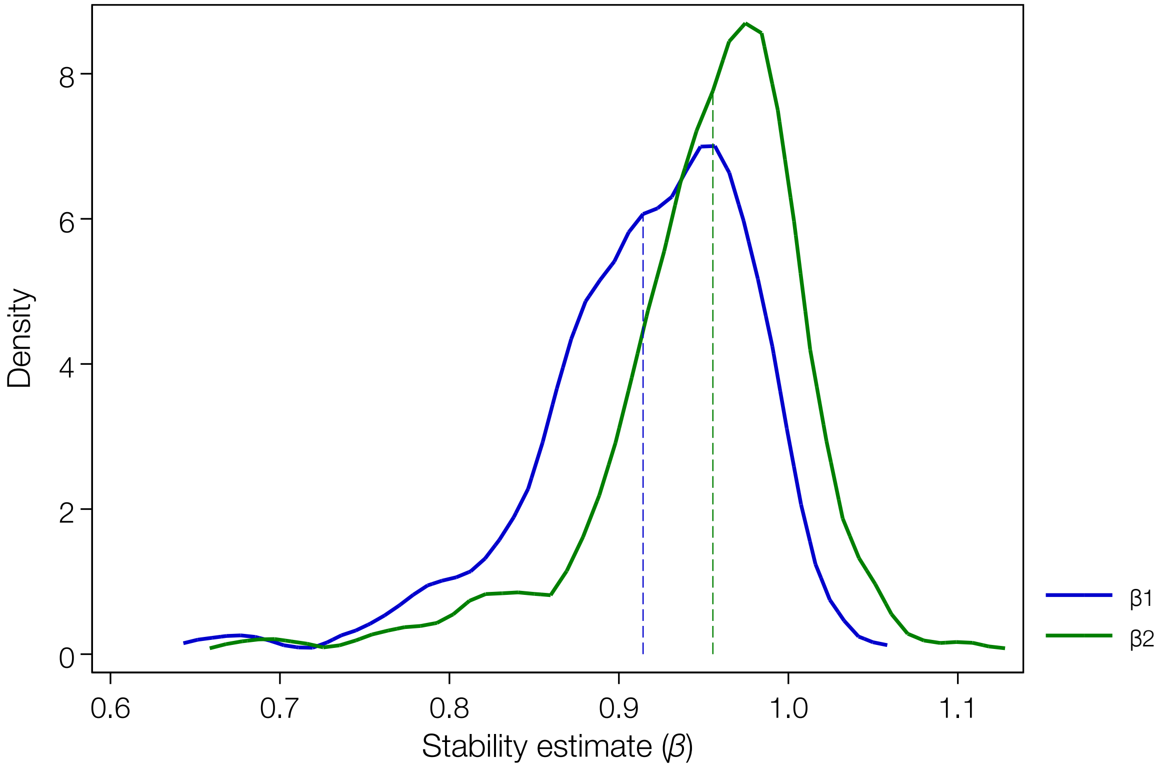

Figure 9: Kernel density curves for stability estimates between the first and second waves (β_1) and second and third waves (β_2) for nonfixed items.

Notes: The vertical dashed lines show the mean values of β_1 and β_2.
Source: Authors' calculations from the General Social Survey three-wave panels, 2006–2014.

We did not think to classify confidence in banks and finance as a recession item, but it stands out as very unstable as well. Considering the role of the banks and financial crisis in precipitating the Great Recession, the instability seems appropriate.

Conclusions

People report their demographic basics and some core identities like religious affiliation and political party identification very reliably. This is terrific news for social surveys and polls, because we all rely heavily on these items either as outcomes in their own right or as predictors of other outcomes. Measurement error in predictors biases statistical estimates toward zero (although, if the error in control variables is relatively greater than that in the independent variables of greatest interest, bias may work the other way). The uniformly high reliability of fundamental variables means that scholars can focus less on this issue than they would have to if reliability was more variable.

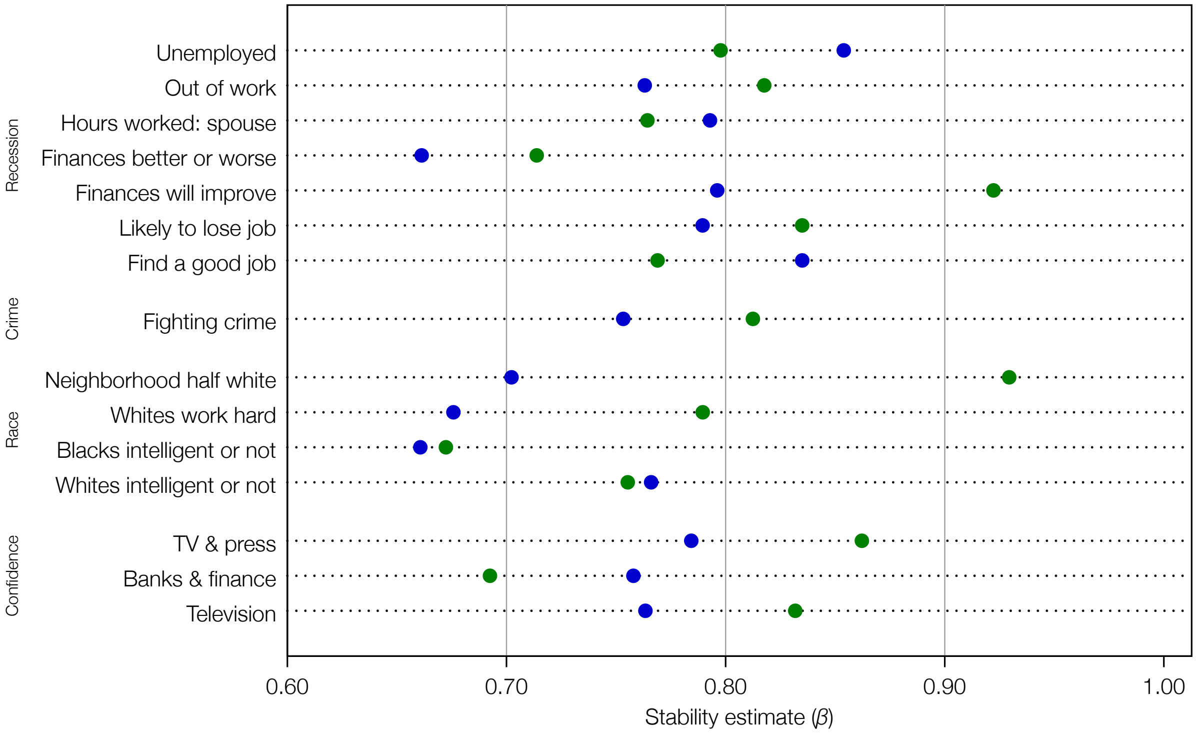

Figure 10: Stability estimates for items with relatively low stability.

Notes: Relatively low stability is defined as either β_1 (blue) or β_2 (green) < 0.80.
Source: Authors' calculations from the General Social Survey three-wave panels, 2006–2014.

People also report familiar beliefs and attitudes quite reliably, though not as reliably as facts and identities. Beliefs about God and life after death are highly reliable, as are attitudes toward issues that are frequently in the news, such as legalizing marijuana, gay marriage, abortion, and capital punishment. Attitudes on less salient subjects such as civil liberties and spanking children are less reliable.

Low reliability is a serious problem for stereotypes about gender and race. These items can make people uncomfortable. Many of them require respondents to keep a counterfactual in mind and then answer. Others stipulate discrimination and ask respondents about its source, putting dissenters who see no discrimination in a quandry. Increasingly, researchers have switched to subliminal and other subtle tools for assessing racial and gender attitudes (Fiske and North 2014). We found that the distinction between the stereotype is true and the stereotype is false was more reliable than trying to ascertain degrees of stereotyping. Yet, even binary versions of the stereotype items were less reliable than most items.

All of our data came from the GSS panels conducted between 2006 and 2014. Because the GSS is an in-person, face-to-face interview that includes show cards as visual aids, the conditions under which we assessed reliability might have been somewhat better than what researchers can achieve with telephone or self-administered surveys.

The GSS panels spanned a momentous period in American history. The simultaneous occurrence of the Great Recession and Obama's historical presidency changed people's views on the economy and at least two racial stereotypes. The models we used characterized the within-person change in different ways. But it is clear from the results that people experienced change in their economic status, their views of major institutions, and their views of African Americans.

The relatively simple models we used to analyze these data have uncovered and quantified general-purpose ideas about item reliability and broad changes in attitudes and behavior. Further analysis of specific items is warranted, but we leave that to subject topic specialists. We close by noting that the GSS panel data yielded immediate returns in new knowledge, and we anticipate more as researchers delve more deeply into its contents.

Notes

[1] Full results are in the online supplement.

[2] The variation comes from people changing their descriptions of their parents' occupations, resulting in the assignment of a different code, and, ultimately, a different pay or credential score. People reported the occupation and job title; they were not asked to estimate the pay of mother's or father's position. The scores attached to the occupations were derived from the American Community Survey.

[3] Note that the horizontal scale in Figure 3 covers a wider range than that in Figure 1; here it goes from 0.25 to 1.00.

[4] Both of these items reflect answering "looking for work" to the employment status question; they differ in who else was included. Persons not in the labor force were missing from calculations involving `unemployed`; all cases were included for the analysis of `outofwork`.

[5] The percentage of Protestants who reported attending services in a phone survey of Ashtabula County, Ohio, was twice what the authors estimated based on the number of cars in church parking lots in that county that same weekend.

[6] The polychoric correlation between confidence in 2006 and 2010 was virtually zero; dividing by a number close to zero resulted in a reliability greater than 8 (an impossible result). The data for 2008–2010–2012 were slightly more plausible. The correlation between confidence in 2008 and 2012 was 0.09; the estimated reliability was 0.14 (very low but not impossible).

[7] The GSS does not disclose the words in the vocabulary quiz.

[8] These are within sampling error of the official unemployment rates in the summer months of the those years.

[9] The interviews are in even-numbered years, but the questions asked about the previous year, thus the odd years mentioned in the text.

References

Alwin, Duane F. 2007. *Margins of Error: A Study of Reliability in Survey Measurement*. New York: Wiley. http://dx.doi.org/10.1002/9780470146316

An, Weihua, and Christopher Winship. 2016. "Causal Inference in Panel Data with Application to Estimating Race-of-Interviewer Effects in the General Social Survey." *Sociological Methods and Research*. forthcoming. http://dx.doi.org/10.1177/0049124115600614

Bielby, William T., Robert M. Hauser, and David L. Featherman. 1977a. "Response Errors of Nonblack Males in Models of the Stratification Process." *Journal of the American Statistical Association* 72 (360a): 723-735. http://dx.doi.org/10.1080/01621459.1977.10479948

Bielby, William T., Robert M. Hauser, and David L. Featherman. 1977b. "Response Errors of Black and Nonblack Males in Models of the Intergenerational Transmission of Socioeconomic Status" *American Journal of Sociology* 82: 1242-1288. http://dx.doi.org/10.1086/226465

Brooks, Clem, and Jeff Manza. 2013. "A Broken Public: Americans Responses to the Great Recession." *American Sociological Review* 78: 727-748. http://dx.doi.org/10.1177/0003122413498255

Clogg, Clifford C. and Darwin Sawyer. 1981. "A Comparison of Alternative Models for Analyzing the Scalability of Response Patterns." *Sociological Methodology* 12: 240-280. http://dx.doi.org/10.2307/270743

Cotter, David A., Joan M. Hermsen, and Reeve Vanneman. 2011. "The End of the Gender Revolution? Gender Role Attitudes from 1977 to 2008." *American Journal of Sociology* 117: 259-289. http://dx.doi.org/10.1086/658853

Davis, James A., and Tom W. Smith. 1980. *The General Social Surveys, 1972-1980: Cumulative Codebook.* Chicago: NORC.

Duncan, Otis Dudley, Magnus Stenbeck, and Charles J. Brody. 1988. "Discovering Heterogeneity: Continuous Versus Discrete Latent Variables." *American Journal of Sociology* 93: 1305-1321. http://dx.doi.org/10.1086/228902

Fiske, Susan T., and Michael S. North. 2014. "Measures of Stereotyping and Prejudice: Barometers of Bias." Chapter 24 in Gregory J. Boyle, Donald H. Saklofske, and Gerald Matthews (Eds.), *Measures of Personality and Social Psychology Constructs*. Elsevier/Academic Press. http://dx.doi.org/10.1016/B978-0-12-386915-9.00024-3

Gelman, Andrew, and Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel / Hierarchical Models.* New York: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511790942

Green, Donald, Bradley Palmquist, and Eric Schickler. 2002. *Partisan Hearts and Minds: Political Parties and the Social Identities of Voters.* New Haven: Yale University Press.

Grusky, David B., Bruce Western, and Christopher Wimer. 2011. "The Consequences of the Great Recession." Chapter 1 in *The Great Recession*, edited by David B. Grusky, Bruce Western, and Christopher Wimer. New York: Russell Sage Foundation.

Hadaway, C. Kirk , Penny Long Marler, and Mark Chaves. 1993. "What the Polls Don't Show: A Closer Look at U.S. Church Attendance." *American Sociological Review* 58: 741-752. http://dx.doi.org/10.2307/2095948

Halpern-Manners, Andrew, John Robert Warren, and Florencia Torche. 2016. "Panel Conditioning in the General Social Survey." Sociological Methodology 45: forthcoming. http://dx.doi.org/10.1177/0049124114532445

Hauser, Robert M., and J. Robert Warren. 1997. "Socioeconomic indexes for occupations: A review, update and critique." *Sociological Methodology* 27: 177-298. http://dx.doi.org/10.1111/1467-9531.271028

Heise, David R., 1969. "Separating Reliability and Stability in Test-Retest Correlation." *American Sociological Review* 34: 93-101. http://dx.doi.org/10.2307/2092790

Hout, Michael. 1999. "Abortion Politics in the United States, 1972-1994: From Single Issue to Ideology." *Gender Issues* 18: 3-34. http://dx.doi.org/10.1007/s12147-999-0013-9

Hout, Michael. 2004. "Getting the Most Out of the GSS Income Measures." GSS Methodological Report 101. Chicago: NORC. http://publicdata.norc.org:41000/gss/documents//MTRT/MR101

Hout, Michael, and Claude S. Fischer. 2014. "Explaining Why More Americans Have No Religious Preference: Political Backlash and Generational Succession, 1987-2012." *Sociological Science* 1: 423-447. http://dx.doi.org/10.15195/v1.a24

Hout, Michael, and Orestes P. Hastings. 2012. "Reliability and Stability Estimates for the GSS Core Items from the Three-wave Panels, 2006–2010." GSS Methodological Report 119. Chicago: NORC. http://publicdata.norc.org:41000/gss/documents//MTRT/MR119.

Hout, Michael, and Orestes P. Hastings. 2014. "The Social Recession: How Americans Have Responded to the Great Recession and Slow Recovery, 2006–2012." Paper presented at the annual meeting of the Population Association of America, Boston, 1 May 2014.

Hout, Michael, Tom W. Smith, and Peter V. Marsden. 2015. "Prestige and Socioeconomic Scores for the 2010 Census Codes." Methodological Report MR124, Chicago, NORC. http://gss.norc.org/get-documentation/methodological-reports

Malhotra, Neil, and Jon A. Krosnick. 2007. "Psychometric Properties of the GSS Wordsum Vocabulary Test." GSS Methodological Report 111. Chicago: NORC. http://publicdata.norc.org:41000/gss/documents//MTRT/MR111.

Owens, Lindsay A., and Karen S. Cook. 2013. "The Effects of Local Economic Conditions on Confidence in Key Institutions and Interpersonal Trust after the Great Recession." *Annals of the American Academy of Political and Social Science* 650(1): 274-298. http://dx.doi.org/10.1177/0002716213500636

Owens, Lindsay A., and David Pedulla. 2014. "Material Welfare and Changing Political Preferences: The Case of Support for Redistributive Social Policies." *Social Forces* 92: 1087-1113. http://dx.doi.org/10.1093/sf/sot101

Silver, Brian, and Barbara Anderson. 1986. "Measurement and Mismeasurement of the Validity of the Self-Reported Vote." *American Journal of Political Science* 30: 771-785. http://dx.doi.org/10.2307/2111272

Smith, Tom W. 1997. "Factors Relating to Misanthropy in Contemporary American Society." *Social Science Research* 26: 179-196. http://dx.doi.org/10.1006/ssre.1997.0592

Smith, Tom W., Peter V. Marden, Michael Hout, and Jibum Kim. 2015. *The General Social Surveys, 1972–2014: Cumulative Codebook.* Chicago: NORC. http://www3.norc.org/GSS+Website/Documentation/.

Steensland, Brian, Jerry Z. Park, Mark D. Regnerus, Lynn D. Robinson, W. Bradford Wilcox, and Robert D. Woodberry. 2000. "The Measure of American Religion: Toward Improving the State of the Art." *Social Forces* 79: 291-318. http://dx.doi.org/10.2307/2675572

Stouffer, Samuel A. 1955. *Communism, Conformity, and Civil Liberties.* New York: Wiley.

Treiman, Donald T. 2009. *Quantitative Data Analysis.* San Francisco: Jossey-Bass.

Wiley, David, and James Wiley. 1970. "The Estimation of Measurement Error in Panel Data." *American Sociological Review* 35: 112-117. http://dx.doi.org/10.2307/2093858

Acknowledgements: The National Science Foundation supported this research through its core grant to NORC for the Data Program for the Social Sciences (SES-12-60440); Hout received supplemental research support from New York University. We benefitted from comments of colleagues, especially Kenneth Bollen, Mark Chaves, the late James A. Davis, Andrew Gellman, Nora Cate Schaeffer, Judith Seltzer, Tom W. Smith, and Christopher Winship. Errors that remain are, of course, our own. Views expressed in the paper do not necessarily represent the views of the funders.

Michael Hout: Department of Sociology, New York University.
E-mail: mikehout@nyu.edu.

Orestes P. Hastings: Department of Sociology, University of California, Berkeley.
E-mail: ophastings@berkeley.edu.