UNIVERSITY OF CALIFORNIA, BERKELEY

2420 BOWDITCH STREET BERKELEY, CA 94720-5670

21 June 2012

Reliability and Stability Estimates for the GSS Core Items from the Three-wave Panels, 2006–2010

Michael Hout & Orestes P. Hastings
University of California, Berkeley

GSS Methodological Report 119

Note:
A coding error led to incorrect estimates for the civil liberties scale items.
This version corrects those errors.

August 2014

ABSTRACT. We assess the reliability and stability of core items in the General Social Survey using Alwin's (2007) implementation of Heise's (1969) model. Of 265 core items examined we find mostly positive results. Eighty items (over 30 percent) have reliability coefficients greater than 0.85; another 84 (32 percent) have reliability coefficients between 0.70 and 0.85. Facts are generally more reliable than other items. Stability was slightly higher, overall, in the 2008-2010 period than the 2006-2008 period. The economic recession of 2007-09 and the election of Barack Obama in 2008 altered the social context in ways that may have contributed to instability.

Reliability and Stability Estimates for the GSS Core Items from the Three-wave Panels, 2006–2010

Introduction

The General Social Survey (GSS) is among the most commonly used data sets in social sciences. The facts, attitudes, values, and opinions collected from representative American households are the primary source of information about almost 200 "core" items; other items in the core that are available from other sources help anchor the unique information.

Assessing the quality of the data is an important part of the project. We know a great deal about the quality of the GSS sample (e.g., Smith, Marsden, and Hout 2011, Appendix A) but 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 questions. Many core items replicate questions asked in other surveys (e.g., Davis and Smith 1980, appendix N). All items are pretested and vetted through cognitive interviews before they enter the survey. Researchers frequently aggregate items into scales and assess the reliability of the constructed scale by modeling item-to-item variation (e.g., Clogg and Sawyer 1981; Treiman 2007, Ch. 9).

With the advent of the GSS panel, conducted in three waves at two-year intervals beginning in 2006, we can use a simple, powerful latent-variable model to estimate reliabilities for most variables in the GSS core (Heise 1969; Wiley and Wiley 1970; Alwin 2007). In the wholesale modeling of most variables in a large survey, we follow Alwin (2007) who used the same model to estimate reliabilities for the American National Election Study panels of the 1950s, 1970s, and 1990s, plus three other three-wave panel studies.

The goal is a broad-brush assessment of question quality. 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 the items and the substantive issues of interest can undo overgeneralizing, sometimes downgrading a positive initial assessment and sometimes upgrading a negative initial assessment. We trust that the research community will undertake this kind of close examination in the coming months. But as an initial foray into the quality of GSS core items, it is more important to get basic estimates of the comparative reliability of all the variables.

We present the model in the next section. In subsequent sections we discuss some decisions we had to make about specific questions, classify the variables into types and subtypes, present the main results by variable, subtype, and broad type, and briefly focus on a few items that appear to violate the assumptions of the model. We conclude by proposing an agenda for future research on the quality of GSS variables.

Models of Reliability and Stability

Heise (1969) proposed the three-variable path model in Figure 1 for what he referred to as "test-retest correlations." The model says that in each wave t of the panel, the observed value on the variable of interest for person i, y_{it}, is the sum of a true score, Y_{it}, and measurement error, ϵ_{it}, that is uncorrelated with Y_{it}. True scores may change between wave t and wave t' in response to "instability" in Y, $\beta_{t't} < 1$, or the influence of exogenous factors, $Z_{it'}$, that are uncorrelated with the previous value, Y_{it}. The observation $y_{it'}$ can differ from a previous observation y_{it} because the true score $Y_{it'}$ differs from the previous one Y_{it} or new errors occurred $\epsilon_{it'}$. All these propositions are implied by these five equations:

$$\begin{aligned} y_{i0} &= \lambda_0 Y_{i0} + \zeta_0 \epsilon_{i0} \\ y_{i1} &= \lambda_1 Y_{i1} + \zeta_1 \epsilon_{i1} \\ y_{i2} &= \lambda_2 Y_{i1} + \zeta_2 \epsilon_{i2} \\ Y_{i1} &= \beta_{10} Y_{i0} + Z_{i1} \\ Y_{i2} &= \beta_{21} Y_{i0} + \beta_{20} Y_{i1} + Z_{i2} \end{aligned} \tag{1}$$

where $i = 1, \ldots, N$, and the λs, ζs, and βs are unknown parameter values. The model embeds a number of key assumptions about the unobserved variables ϵ and Z. Specifically they are uncorrelated with the observed variables y, the true scores Y, and one another. They are quite reasonable theoretical assumptions that keep measures and true scores distinct. Even with these simplifications, the model is too complicated; it has nine parameters to be estimated but we have just the three covariances among the three observed measures to work with.

The defining simplification is to assume that there is no lagged effect of Y_0 on Y_2, i.e., β_{20} = 0. Another key assumption, introduced by Heise (1969), reduces the three λs and three ζs

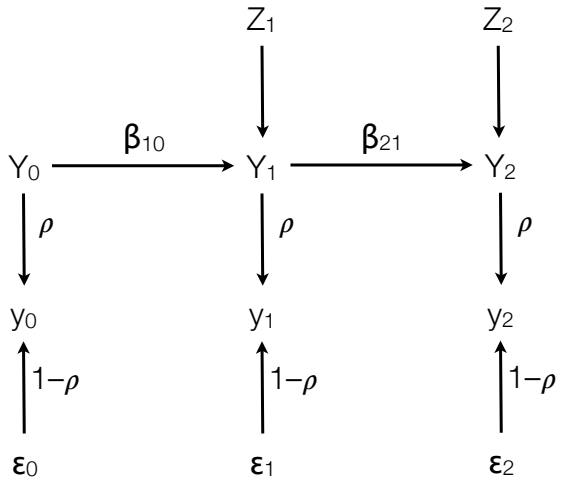


Figure 1. Heise's Model of the Reliability and Stability in a Three-wave Panel Study
Source: Authors' redrawing of Figure 5 in Heise (1969).

to a single parameter, called ρ. First ρ links each λ to the corresponding ζ:

$$\begin{aligned} \lambda_t &= \rho_t \\ \zeta_t &= 1 - \rho_t \quad . \end{aligned} \tag{2}$$

Linking each λ and ζ in this way constrains the variances of the two latent variables Y and ϵ to always add up to the observed variance in y — a reasonable constraint that helps define the two latent variables. Second, Heise (1969) proposes specifying that the ratio of true to observed variance is the same in each panel. This implies that ρ does not change over time, i.e., $\rho_0^2 = \rho_1^2 = \rho_2^2 = \rho^2$, and ρ^2. With all this in hand, we can estimate ρ^2 as a simple function of the correlations among the three observed ys:

$$\rho^2 = \frac{r_{01} r_{12}}{r_{02}} \tag{3}$$

where $r_{tt'}$ for $t' \neq t$ can be either a Pearson correlation or a polychoric correlation (Alwin 2007, pp. 84-86). As the model is just-identified, the βs also have simple formulas:

$$\beta_{10} = \frac{r_{01}}{\rho^2} \text{ and } \beta_{21} = \frac{r_{12}}{\rho^2} \, . \tag{4}$$

Other models and interpretations exist (Wiley and Wiley 1970; Alwin 2007). For example, if ρ^2 varies from panel-to-panel, then $\rho_1^2 = \frac{r_{01} r_{12}}{r_{02}}$, i.e., what we will characterize here as the

constant reliability of y is actually its reliability in the middle wave of the three-wave panel. Alwin (2007, p. 107) argues that the Wiley-Wiley model only makes sense where one has kind of dynamic equilibrium such that variances do not change across waves. If variation either increases or decreases over time, then the Wiley-Wiley model mistakes that for falling or rising reliability. Regardless, reliability estimated at wave 2 under Wiley-Wiley equals the single estimated reliability in the Heise (1969) model (Alwin 2007, p. 107).

Modifications and Scales

The models described in the preceding section apply when the observed variables are dichotomous, ordered, or continuous variables. Many GSS core variables have three or more unordered categories. For widely used variables with relatively few categories — marital status, employment status, and religions — we made some dichotomies out of the categories and used the Heise model to estimate reliability and stability. Others — notably ethnicity and ancestry — proved to be harder to reduce.

Table 1. Dichotomies Formed from Categorical Variables

Categorical variable	GSS mnemonic	Dichotomy	New mnemonic
Marital status	marital	Married vs. other	Married
		Never married vs. other	Nevermar
Employment status	wrkstat	Employed vs. other	Employed
		Unemployed vs. other labor force	Unemployed
		Retired vs. other	Retired
Current religion	relig	No religion vs. some	None
Religion raised in	relig16	No religion vs. some	None16

Some categorical variables — occupations, for example — have widely accepted scores, even though the unordered categories themselves cannot be analyzed with the Heise model. For occupations, we assess their reliability by looking at the reliability of prestige and SEI scores (Hauser and Warren 1997). We also made use of the "Reltrad" recode of detailed religious denominations developed by Steensland et al (1999) to construct additional religion dummy variables — "TradEvang," "TradMain," "TradBlack," "TradCath," "TradJew," "TradOther," and "TradNone," for conservative Protestants, mainline Protestants, African-

American Protestants, Catholics, Jews, other religions, and no religion, respectively. Because these dummy variables exhaust the information in `Reltrad`, they are not independent of one another. Knowing the reliabilities of six of these items would allow a researcher to derive the seventh. We use the same approach to assess the reliability of religious origins and spouses' religions.

We also constructed three widely used scales for vocabulary, support for legal abortion, and gender roles. The GSS vocabulary quiz has ten words (Malhotra and Krosnick 2007). Rossi's original abortion attitudes scale uses six questions; asking about abortion under any circumstances (`abany`) is a common extension (e.g., Hout 1999); we estimate the reliability of both scales. Finally, four gender-typing items are often used to make a scale (Cotter, Hermsen, and Vanneman 2011).

We supplement these common scales with additional ones devised for this study. We combine five questions about suicide and end-of-life to form a suicide scale. We combine Stouffer's (1955) civil liberties items into four scales regarding the freedom of atheists, communists, 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 socializing scale.

Exclusions

We excluded the objective geographical measures — sampcode, srcbelt, size, xnorcsiz. NORC codes them from address of the interview; they are part of the administrative record and not responses to questions the GSS poses to either the respondent or the interviewer. We also excluded aspects of the interview such as the number of people enumerated in the household, their ages, the respondent's relation to the householder, whether the respondent was permanent resident or visitor in the household, and whether the interviewer was hispanic because they were not constrained to be the same across waves of the panel; in short, we thought the model in Figure 1 to be inappropriate for these variables.

Aggregating 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 so we modified his scheme for our purposes. We subdivided facts into demographic, socioeconomic, religious, and behavioral facts. The words of the vocabulary quiz do not fit neatly into any of the other categories, so we made "words" a new type. We combined beliefs and values and then subdivided them by topic: sex-and-sexuality, religious, social, civil liberties, gender-and-family, and racial. The GSS asks respondents to render descriptions of others as well as themselves, so we generalized the self-description type into a category of self and other descriptions that we call "placements and evaluations." We subdivided attitudes into those that address institutions, civil liberties, and other socio-political issues. These five types and sixteen subtypes are heuristics we use to help us organize and discuss the reliability and stability estimates. The aggregation does not affect the calculations with respect to individual questions. Of course the averages for the types and subtypes would be different if we reclassified some items.

We also distinguish fixed items from those that can change. Fixed items should, by definition, have stability coefficients of 1.0 plus or minus sampling error. We could use that information to constrain the estimated reliability or to even estimate more than one reliability coefficient per item. We have done neither. Instead we have estimated the stability coefficients as a test of Heise's model in this setting. If any stability coefficient for a fixed item differs significantly from 1.0, then we have to figure out why. Among the possibilities is the prospect Figure 1 is the wrong model. With no degrees of freedom for testing the model elsewhere, we regard this test as important.

An alternative model for fixed items and some less time-sensitive traits like vocabulary would specify a single latent Y that does not change; we thank Steve Vaisey (personal communication) for this suggestion. Reliability under this model could be assessed using iterative structural equation modeling (SEM) methods. Its beyond the scope of this paper, but an interesting alternative worth exploring in future research.

Table 2. Reliability Summary Statistics by Type and Subtype of Item

Type of item	Subtype of item	Median	Mean	Standard deviation	Number of items
Facts	**All**	**.918**	**.882**	**.122**	**88**
	Demographic	.958	.933	.110	22
	Religious	.964	.940	.060	21
	SES: fixed	.887	.880	.065	10
	SES: can change	.814	.839	.132	25
	Behaviors	.754	.757	.147	10
Words	**All**	**.750**	**.782**	**.163**	**11**
Beliefs and values	**All**	**.706**	**.690**	**.151**	**97**
	Sex & sexuality	.839	.801	.093	16
	Religious	.811	.811	.096	12
	Other social	.757	.711	.127	20
	Civil liberties	.709	.720	.084	20
	Gender & family	.622	.594	.098	12
	Racial	.490	.515	.172	16
Placements & evaluations	**All**	**.675**	**.673**	**.127**	**22**
Attitudes	**All**	**.658**	**.664**	**.127**	**63**
	Institutional	.600	.598	.065	14
	Taxes & spending	.681	.663	.121	29
	Other social & political	.710	.710	.150	20

Source: Authors' calculations from General Social Surveys, 3-wave panel, 2006-2010.

Results I: Reliability Patterns

The main results are in Table 2 and Figures 2A-C. The table shows descriptive statistics on the reliability of each type and subtype of item. The figures show the reliability and stability estimates for each item as well as the three correlations — r_{01}, r_{12}, and r_{02} that determine reliability and stability; we use the polychoric correlations for items with less than 11 possible values and Pearson correlations for items with 11 or more possible values (see Alwin 2007 for rationale). The items are arrayed in the figures in the order of descending average cross-panel correlation (i.e., $(r_{01} + r_{02} + r_{12})/3$).

Facts are the most reliable type of item, followed, in order, by words, beliefs, placements, and attitudes. Five items have reliability estimates closer to 1.1 than to the theoretical maximum of 1.0.[1] It seems reasonable to think that for these five items reliability is very high but

[1] `confed`, not included in the table or figures, had a reliability coefficient estimated to be 10.0. We discuss that improbable estimate in the next section.

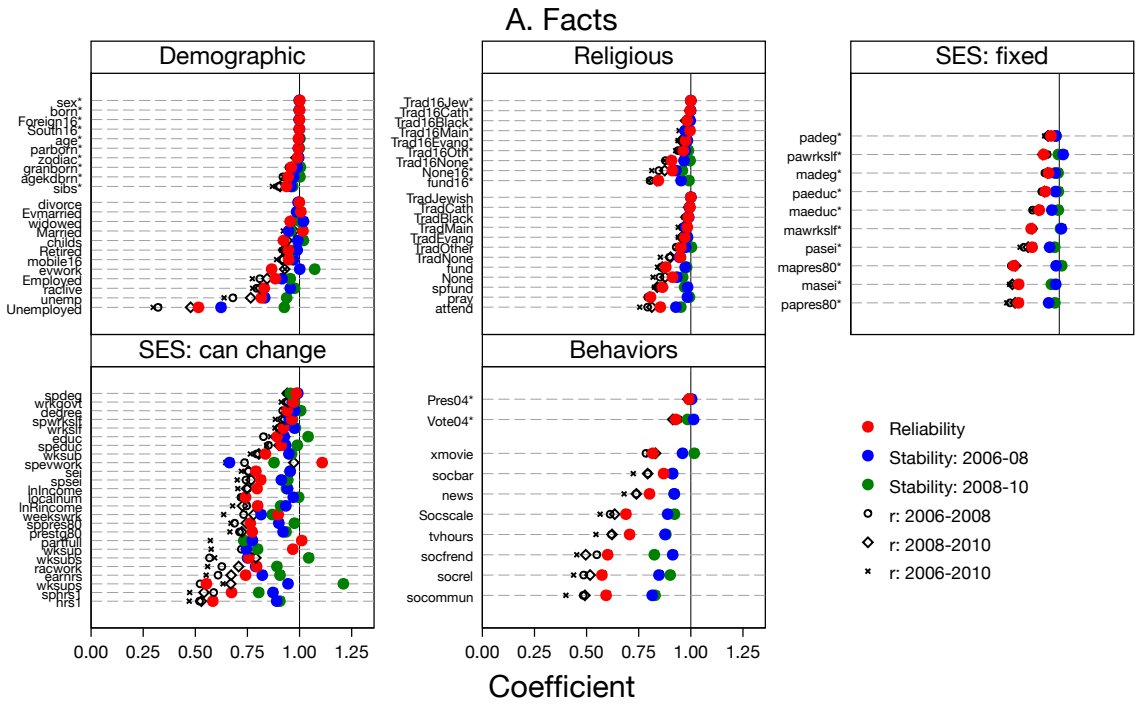
greater than 1.0 only because the correlations, each subject to some sampling error, implied it by chance.[2]

Demographic and religious facts are particularly reliable; their reliability estimates average 0.96 and 0.94, respectively. These include contemporaneous facts like age, education, employment status, marital status, and religious denomination as well as facts about the past such as month of birth (recoded to sign of the zodiac and treated as a score), age at the birth of first child, number of siblings, and religion raised in. A number of these are fixed attributes. In each case of a fixed attribute in these categories, the stability estimates are close to 1.0.

The lowest reliabilities among demographic facts concern the labor force, and we suspect that the recession played a role in low reliability estimates for `Employed`, `Unemployed`, and `unemp`. The Heise-Alwin model makes a Markov-like assumption that initial statuses do not affect status in the last panel, except indirectly through effects on the middle panel. That might not be true with respect to labor force status during a recession. For example, we suspect that the probability of moving from non-employment to employment between 2008 and 2010 might be higher for those employed in 2006 than those not employed in 2006. If so then the model's assumption that $\beta_{20} = 0$ is violated, and we cannot separate stability from reliability without replacing the Markov-like assumption of no lagged effect with another assumption that restricts parameter values.

Socioeconomic facts are only slightly less reliable than demographic and religious facts. Education and occupation, as well as spouse's and parents' educations and occupations, are almost as reliable as demographic and religious items, ranging from 0.75 to 0.90. The equivalence between self-reports and (retrospective) reports about others is a common, if surprising, finding (e.g., Bielby, Hauser, and Featherman 1977a 1977b; Alwin 2007). Our estimates are right in the middle of the range of previous estimates (Alwin 2007, pp. 302-308). Income, both personal earnings and total family income, are reported slightly less reliably (and with substantially more missing data) than education and occupation. Hours worked last week, both personal and spouse's, are the least reliably reported socioeconomic facts. As with employment status indicators, we suspect that the Markov-like assumption that initial work hours do not affect last-wave work hours might not apply for those whose hours in 2008 were reduced

[2]Unfortunately, because our estimation strategy relies on the ratio of correlations, $\rho_1^2 = \frac{r_{01}r_{12}}{r_{02}}$, many of which are polychoric, we have no estimates of the standard errors with which to test our conjecture.



Note: Variables sorted from highest to lowest average correlation. Fixed variables are starred.

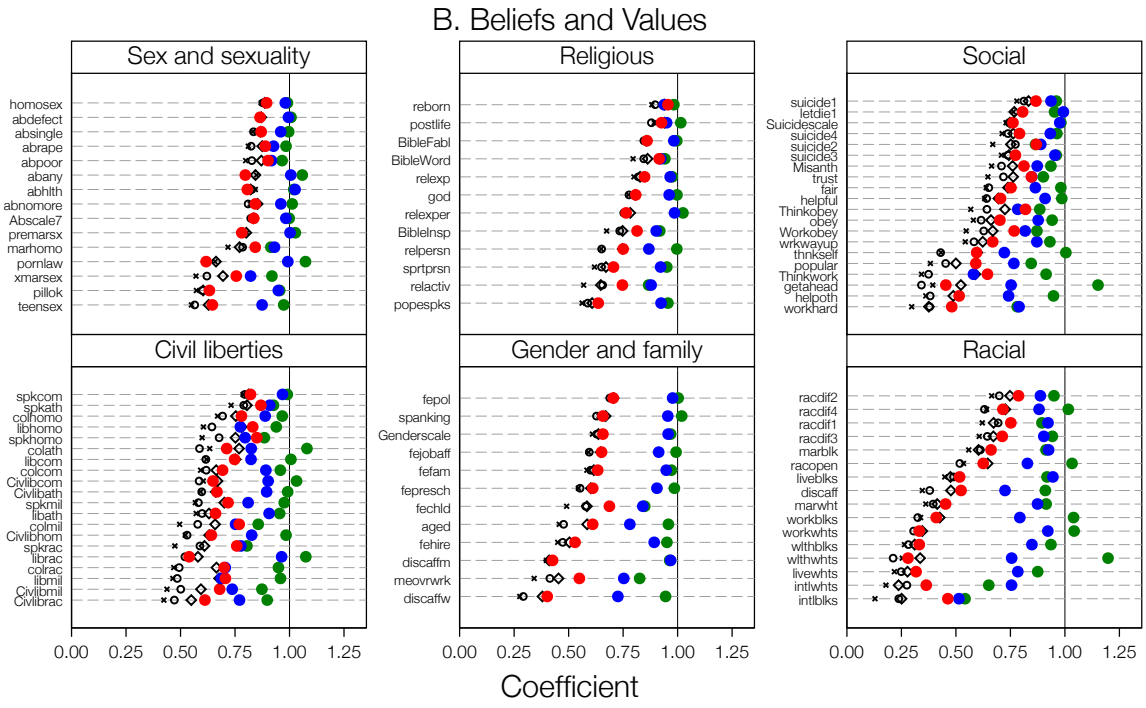
Figure 2A. Reliability, Stability, and Correlations by Subtype of Item: Facts
Source: Authors' tabulations General Social Survey three-wave panels, 2006-2010.

due the economic recession.

Reports about voting behavior — whether the respondent voted in the 2004 presidential election and, if so, for Bush or Kerry — are very reliable, but the other behavioral facts are substantially less so. Socializing with friends, neighbors, and relatives are particularly unreliable compared with other facts; estimates are about 0.60 for all three.

Some of the words of the vocabulary quiz are highly reliable but three are not very reliable at all. The `Wordsum` scale is right in the middle of the (wide) range of single-item reliabilities. The variation is not related to the difficulty of the words. The most reliable words (`Wordb` and `Worde`) are among the easiest words (90 and 79 percent of respondents, respectively, give the correct definition), but the third and fourth most reliable are the hardest and third hardest words (`Wordc`, 21 percent correct; `Wordh`, 34 percent correct). Malhotra and Krosnick (2007) discuss other issues with reliability of the GSS words. We note that their proposed four-item alternative vocabulary scales are substantially less reliable than `Wordsum` itself.[3]

[3]Their alternative scale A (words a, c, h, and i) has a reliability of .58, their alternative scale B (words d, e, f, and g) has a reliability of .65, an alternative composed of the four most reliable words in our estimation (words



Note: Variables sorted from highest to lowest average correlation. Fixed variables are starred. See Figure 2A for key.

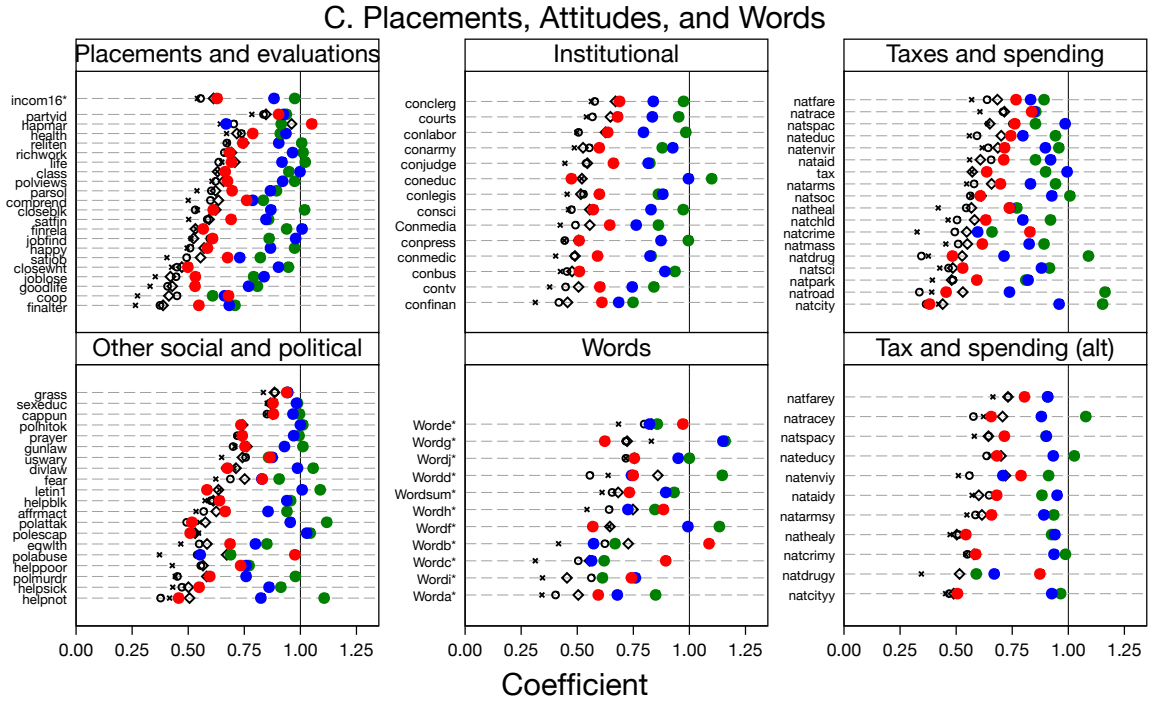
Figure 2B. Reliability, Stability, and Correlations by Subtype of Item: Beliefs and Values
Source: Authors' tabulations General Social Survey three-wave panels, 2006-2010.

Beliefs and values are the third most reliable type of item, on average. Again there is substantial variation across items in this type. Beliefs and values having to do with sex and sexuality are the most reliable in this type. Reliability estimates for questions about abortion and various forms of adult sexual behavior range from 0.76 to 0.90. Answers to questions about laws and teen sex are less reliable (between 0.62 and 0.65).

Religious beliefs and values are also relatively reliable for subjective items. Belief in God and the afterlife as well as the truth of the Bible all have reliabilities great than 0.75. Two of the three reports about religious experiences are also that high. Identity and activity items are slightly less reliable but still over 0.65.

Social beliefs and values about the end of life and social trust are relatively reliable — all above 0.70. Beliefs about the relative importance of five traits in children are much less so. But the key distinction — whether it is more important for children to think for themselves or be obedient — has a very good reliability estimate of 0.82. The least reliable item in this subtype is `get ahead` — whether hard work or luck is more important to success — which

b, c, e, and h) has a reliability of .70, and `Wordsum` has reliability of .73.



Note: Variables sorted from highest to lowest average correlation. Fixed variables are starred. See Figure 2A for key.

Figure 2C. Reliability, Stability, and Correlations by Subtype of Item: Placements, Attitudes, and Words

Source: Authors' tabulations General Social Survey three-wave panels, 2006-2010.

has an estimated reliability of only 0.45.

The famous Stouffer civil liberties items have an average reliability of 0.72. The six questions Stouffer (1955) included in his original analysis — the atheist and communist items – are not noticeably more reliable than the others. The only item with a reliability below 0.65 is the one that asks about removing a racist book from the library. This is also the item with the most missing data; people seem to have a fewer fixed ideas on this point than on the others in this subtype.

The belief and values items that refer to racial and gender differences are far less reliable than other beliefs and values. These are very important items for GSS users. Their low reliability is a matter of great concern. The 2008 election cycle was historic not only in its result — the election of the first African American president — but for then-Senator Hilary Clinton's competitiveness in the Democratic primaries and then-Governor Sarah Palin's campaign as the Republican Vice Presidential candidate. These unprecedented candidacies may have changed the meaning of questions about race and gender for some respondents.

Table 3. Confidence in People Running Financial Institutions by Wave

		Weighted Count		
		2010		
2006	2008	A great deal	Only some	Hardly any
A great deal	A great deal	29.3	54.2	14.0
	Only some	8.3	77.5	46.6
	Hardly any	1.6	6.7	13.0
Only some	A great deal	5.6	52.3	10.6
	Only some	10.9	172.4	108.3
	Hardly any	3.1	26.8	69.5
Hardly any	A great deal	0	1.5	4.8
	Only some	1.2	24.0	26.1
	Hardly any	1.0	7.8	47.0

Source: Authors' calculations from General Social Surveys, 3-wave panel, 2006-2010.

Results II: Stability Patterns

The main interest in panel data, is, of course, the prospect of uncovering individual change. That change, apart from erroneous differences that occur due to unreliability, comes in two forms. The first is change in the rank order of true scores — usually called "instability" — and measures as the departure of the $\hat{\beta}$ coefficients from 1.0. The second is the shift of the marginal distribution and/or mean of an item, independent of the rank order of persons. This second source of change is generally ignored in the psychometric literature, and it lacks a common name or referent. It resembles what it known in the social mobility literature as "structural mobility." In many applications, this "structural change" in the marginal distribution will, in fact, be more interesting than either unreliability or instability.

To help fix ideas, consider the example of confidence in financial institutions. In each wave of the panel, people who got either ballot B or ballot C of the GSS were asked whether they had "a great deal," "only some," or "hardly any" confidence in the "people running" many institutions, including banks and financial institutions. Among the 821 people who gave valid answers to the question in all three surveys, the fraction expressing hardly any confidence rose from 15 percent in 2006 to 23 percent in 2008 to 42 percent in 2010. Clearly structural change was a factor; the distribution moved from greater to lesser confidence. We cannot infer from the marginal shift, however, that the answers were unstable. If predicting the answers in 2010

from answers in 2008 was as easy or hard as predicting answers in 2008 from answers in 2006, then we would conclude that confidence — relative to the shifting context — was stable. The average of the $\hat{\beta}$ coefficients for `confinan` is 0.72. Though substantially less than 1.0, it indicates that the overall decline in confidence was accompanied by only modest reversals in the rank order of individuals from most to least confident.[4]

We actually measure the complement of change in the stability parameters of the Heise-Alwin model. No change net of unreliability yields a stability estimate near 1.0; significant change pushes the stability estimate toward zero. Stability, in this framework, is relative to the marginal distribution of the variable at the initial and times. Across-the-board increases that shift everyone up or down do not reduce our ability to predict answers and thus do not reduce stability estimates. On the other hand, changes that vary from person to person make it difficult to predict a person's response in one year from what she said before, yielding low stability estimates. The Heise-Alwin model yields two coefficients for each item; one for the transition from the initial to the middle year and the other for the transition from the middle to the final year. Figures 2A-2C show them in blue (2006-2008) and green (2008-2010).[5]

Before discussing the results for types and subtypes of items it is important to put the panel waves in the context of events at the time. The Great Recession started just before fieldwork for the 2008 GSS began; the recession began in December 2007 and interviews began in April 2008. Surprisingly, people were unaware of the recession at first. Unemployment was still just 5.0 percent as interviewing began and 6.1 percent in September as the last interviews concluded. It took the NBER recession dating committee until December 2008 to announce that the economy had been in recession since the previous December. According to "Google trends," mentions of "recession" spiked in January 2008 then abated until the transition period between the presidential election in November 2008 and the inauguration in January 2009.

As we discussed above, the disruptions of the recession might invalidate the Heise-Alwin model for some variables that are particularly indicative of the recession. We noted already

[4]It is worth noting that floor and ceiling effects could be substantial with an item like this as it has only three response options.

[5]We can also show that stability is the ratio of the two-wave correlation to what might be thought of as the off-year correlation, i.e., $\beta_{10} = r_{02}/r_{12}$ and $\beta_{21} = r_{02}/r_{01}$. That implies that if stability is perfect over one period, then the two-wave correlation will equal the correlation over the other interval, i.e., $r_{02} = r_{12} if \beta_{10} = 1$ and $r_{02} = r_{01} if \beta_{21} = 1$.

that we suspect that employment status and work hours in 2006 may directly affect these statuses in 2010, net of status in 2008. And thus we get unrealistic estimates of reliability and stability due to violations of the model's assumptions.

A political item makes prospect even clearer. We have excluded `confed` from the discussion to this point because 2010 answers are virtually uncorrelated with 2006 answers to the question "As far as the people running these institutions are concerned, would you say you have a great deal of confidence, only some confidence, or hardly any confidence at all in (e) the executive branch of the federal government?" Of course that is because the 2008 election dramatically changed who was running the executive branch of the federal government. With that in mind, it is slightly surprising that the 2006 and 2010 answers were merely uncorrelated; they could well have been negatively correlated (which would have resulted in a negative reliability estimate and at least one negative stability coefficient).

The economy was not the only source of social change, of course. Voters elected the first black president, gay marriage was a persistent and contentious issue, wars continued in Iraq and Afghanistan, and phones seemed to merge with computers. These trends and more were reflected in changing distributions for many GSS items. The stability coefficients reflect these kinds of changes net of factors that moved distributions up and down net of where people started.

Thus we expected to see more recession-related instability in the 2008-2010 wave than in the 2006-2008 wave. Likewise, as the 2008 interviews were done as the 2008 party primary elections were going on and most were done before the nominees were known, political instability may be greater 2008-2010 than 2006-2010 as well. The data defied that expectation. For items that evinced any instability, it was greater (the stability coefficient closer to zero) in the 2006-2008 wave than in the 2008-2010 wave, as evidenced by the preponderance of green circles to the right of blue circles in Figures 2A-2C. The tendency is very widespread; even the words were more stable in the more recent period.

We expect perfect stability for the fixed items, by definition. If we know a person's age at one time we should be able to perfectly predict her age two years later. Similarly her month of birth, gender, and race should be perfectly stable. Things from her past such as where she grew up, how many siblings she had, the religion she was raised in, and facts about her parents like their educations and occupations (anchored to her teenage years) should

be almost perfectly predictable after adjusting for unreliability. Other facts like marital status, religious denomination, attending religious services, and education change slowly enough that we expect near-perfect stability for these items as well. And that is what the data show.

Socioeconomic facts like jobs and income, on the other hand, shift over time, especially during recession times. The data show that, too. Stability estimates for employment status, hours and weeks of work (for respondent and spouse), occupational status, and income (family and personal) are mostly in the range from 0.75 to 0.85. Socioeconomic standing of the respondent's occupation (`sei`) is more stable (0.96 in both waves), in part because previously employed respondents are asked to describe the job they last had and their SEI is that of the job they lost. Working hours and spouse's working hours, after adjusting for relatively low reliability have moderately high stability (0.89 and 0.87, respectively).

Beliefs and values are far more stable than they are reliable, especially as indicated by the 2008-2010 panel. Very few items in Figure 2B have stability estimates below 0.80. The two that do — `intlwhts` and `intlblks` — hint that then Senator Obama's candidacy in 2008 was changing IQ stereotypes. Overall gender and racial beliefs and values held steadier over this period than the observed data suggested because those observed data were unsettled by the low reliability of these items. However, statistical adjustment is a poor substitute for finding more reliable measures. As we have seen with some other items, the context of the times can render the underlying assumptions of the model moot, undoing the prospect of getting good estimates of reliability.

Nine placements have stability estimates (averaged) of less than 0.90. Six of the nine point to changes in the economy during the recession: the respondent's standard of living compared to that of her parents at the same age (`parasol`, $\beta_{10} = 0.87$; $\beta_{21} = 0.89$), financial satisfaction (`satfin`, $\beta_{10} = 0.85$; $\beta_{21} = 0.86$), the prospect of losing one's job (`joblose`, $\beta_{10} = 0.84$; $\beta_{21} = 0.79$), the prospect of improving one's standard of living (`goodlife`, $\beta_{10} = 0.77$; $\beta_{21} = 0.81$), job satisfaction (`satjob`, $\beta_{10} = 0.73$; $\beta_{21} = 0.82$), and whether the person's financial situation has gotten better or worse (`finalter`, $\beta_{10} = 0.68$; $\beta_{21} = 0.71$). The Great Recession was the country's most significant economic crisis in a generation, millions of families were affected, and it seems right that these stability estimates would turn out so low. It suggests that the model is appropriate for data of this kind.

The other three placements that show relatively low stability are how happy one's marriage

is (hapmar, $\beta_{10} = 0.67$; $\beta_{21} = 0.92$), and two assessments of the respondent by the interviewer (coop and comprend).

Attitudes toward institutions and spending changed substantially as well. The three institutions that faced the biggest changes in public confidence were the executive branch (as noted before), the courts, and financial institutions. The least stable spending items were spending on parks, the environment, health care, drug rehabilitation, and crime.

Variation in Reliability and Stability

We can think of many ways in which reliability could vary systematically among respondents. One thought might be, for example, that college graduates give more reliable answers than people with less education do. We leave those kinds of investigations to the future agenda.

We consider one methodological source of variation in reliability: mode of interview. Of the 1,276 cases interviewed three times, 823 (64 percent) were interviewed in-person all three times, 307 were interviewed in-person twice and by phone once, 120 were interviewed in-person once and by phone twice, 25 were interviewed by phone all three times, and we do not know the mode of the first interview for one of the cases. We reestimated the reliability of 54 popular items for the all-in-person subset and for the other respondents. For 33 of the 54 items we examined, the difference between the reliability when the interview was in-person and when it was by phone was less than 0.10 in absolute value, and another 12 were between 0.10 and 0.20. Six of the nine items that appeared to be sensitive to mode were words. As noted before we have only 280 cases for most words so the estimates of their reliability is much more subject to sampling error than other items are. Furthermore three of the words appear more reliable in-person and the other three appear more reliable over the phone.

That leaves three items that appear to be very sensitive to mode. The last spending item — spending on parks and recreation — has a reliability of 0.72 in person but only 0.52 by phone. It comes at the end of a long list, and the fatigue of repeating the response options may affect the respondent or the interviewer, reducing the reliability of the phone response. But the next to last spending item is not different in-person or by phone, so this is a highly speculative reading of one piece of evidence. An item that asks about the value of hard work (word hard) is more reliable on the phone. We have no idea why. Finally the interviewers'

assessments of respondents' cooperativeness were more reliable over the phone than in person. As interviewers changed across waves (but respondents did not) we regard this more about rapport than method.

In sum, our attempt to find a method effect turned up some idiosyncratic differences by method, but no hint that in-person is consistently more reliable than phone interviewing. Of course relying more on the phone would doubtless reduce response rates. But we would not get less reliable answers from the people who did respond were the GSS to switch rely more heavily on phone interviews.

Agenda for Future Work

The first item on the agenda is to figure out appropriate latent class models for the categorical variables like ethnicity, marital status, and so on that are widely used GSS variables but dichotomized or excluded from this study for lack of a good model.

The second item on the agenda is to assess the difficulties in measuring beliefs about gender and race. Two hypotheses come to mind. The first concerns the context. The 2008 GSS was in the field when then-Senators Obama and Clinton were battling for the 2008 Democratic nomination for president. Their candidacies may well have altered public perceptions, not only of particular aspects of race and gender but also the very meaning of terms in some of the GSS items. That kind of reconsideration is not part of the Heise-Alwin model. This conjecture ascribes the low estimated reliability of the affected items to our choice to use an inappropriate model. The alternative hypothesis takes the low reliability at face value and implies finding new ones with which we might better assess trends over time and differences in the cross-section on these issues. Sometimes low reliability means that we are asking people to answer questions they have not thought about. That is hardly the case with gender and racial issues in the United States. These have been among the most extensive social controversies throughout the GSS time series. Thus if the low reliability is right, then we need new items in these crucial domains. We cannot adjudicate between the context and item hypotheses with the data at hand. Choosing must wait until the 2010-2012-2014 three-wave panel is completed. Respondents in all three of those waves encountered the items after Obama became president. If the first hypothesis is correct, attitudes should be coming to a new equilibrium and the

reliability estimates of racial and gender values should be higher.

We also need to improve the vocabulary quiz. A prior methodological report (Malhotra and Krosnick 2007) pointed to some problems, most notably the way the words that were originally moderately difficult have become difficult. Our analysis differs from theirs in several important ways. Nonetheless, we concur in their view that the quiz needs new moderately difficult items. The small number of observations for these ten items makes it hard to reach a firm conclusion, however.

We propose to extend the analysis to some non-core items that were, nonetheless, measured in all three waves. For example items on sexual behavior and intravenous drug use were part of each survey. So too were several questions about the role of science in American life.

We would also like to replicate some of the estimations within subpopulations to see if items work differently for different kinds of respondents. That will almost certainly require pooling data from these three waves with data from the 2008-2010-2012 panel now in the field.

Some items with complicated selections also invite further analysis. For example, take the `joblose` question. Employed people are asked how likely it is that they will lose their job anytime soon. Those who correctly predict a job loss at one wave do not get asked the question if they are unemployed at the next wave. We plan to concatenate the `joblose` and `wrkstat` items to capture this dynamic.

Conclusions

The GSS core items, especially the most used ones, are mostly very reliable survey items. Facts and most beliefs are particularly reliable. This is important because almost every study uses some facts to condition or explain changes or differences in another item. The lack of alternatives for some of the facts and beliefs that the GSS collects make it imperative that they be high quality items and most are.

Low reliability is a serious problem for beliefs about gender and race, especially the recently introduced ones that replaced items that were no longer useful because ninety percent or more of Americans agreed on the answers. The newer items are lower quality and the time series is suffering accordingly.

The panel is doing a very good job of documenting within person change during a momentous period in American history. The simultaneous recession and historical presidency have changed peoples views on the economy and at least two racial stereotypes. The simple models we used to analyze these data have uncovered and quantified these important changes. In that regard a key goal of collecting panel data has yielded immediate returns in new knowledge.

Acknowledgments

We are grateful to the GSS Board of Overseers, especially Judy Seltzer, for helpful comments on previous drafts of this report. We acknowledge financial support from the National Science Foundation (SES-0824618), the Institute for the Study of Societal Issues, and the Berkeley Population Center (NICHD R21 HD056581). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Science Foundation, the National Institute of Child Health and Human Development, the National Institutes of Health, or the University of California.

References

Alwin, Duane F. 2007. *Margins of Error: A Study of Reliability in Survey Measurement*. New York: Wiley.

Bielby, William T., Robert M. Hauser, and David L. Featherman. 1977a. "Response errors for nonblack males in models of the stratification process." *Journal of the American Statistical Association* 72: 723-735.

Bielby, William T., Robert M. Hauser, and David L. Featherman. 1977b. "Response errors of black and non-black males in models of status inheritance and mobility." *American Journal of Sociology* 82: 1242-1288.

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

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

Fischer, Claude S., and Michael Hout. 2006. *Century of Difference: How American Changed Over the Last One Hundred Years.* New York: Russell Sage Foundation.

Hauser, Robert M., and J. Robert Warren. 1997. "Socioeconomic indexes for occupations: A review, update and critique."*Sociological Methodology* 27: 177-298.

Heise, David R., 1969. "Separating Reliability and Stability in Test-Retest Correlation." *American Sociological Review* 34: 93-101.

Hout, Michael. 1999. "Abortion Politics in the United States, 1972-1994: From Single Issue to Ideology." *Gender Issues* 18: 3-34.

Malhotra, Neil, and Jon A. Krosnick. 2007. "Psychometric Properties of the GSS Wordsum Vocabulary Test." GSS Methodological Report 111. Chicago: NORC.

Smith, Tom W. and Jaesok Son. 2011. "An Analysis of Panel Attrition and Panel Change in the 2006-2008 General Social Survey Panel." GSS Methodological Report 117.

Chicago: NORC. [http://publicdata.norc.org:41000/gss/documents/
/MTRT/MR118.pdf].

Smith, Tom W., Peter V. Marden, and Michael Hout. 2011. *The General Social Surveys, 1972-2010: Cumulative Codebook.* Chicago: NORC.

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.

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

Treiman, Donald T. 2007. "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.