New Method

Kevin Kiley

5/20/2019

Basic Models

Our two different models make different predictions in the three-wave panel context. The AUM makes the following prediction for wave 3:

$$E(y_{i3}) = y_{i2} \tag{1}$$

That is, the AUM predicts that a respondent's most recent response is the best available predictor of her next response. If change is persisting (i.e., if our metaphorical traveler wakes up in the same bed she went to sleep in last night), then our best guess is that she will be close to where we saw her last.

The SDM makes the following prediction:

$$E(y_{i3}) = \mu_i \tag{2}$$

Since the best estimate of μ_i is the mean of the respondent's two previous answers, we can rewrite the SDM prediction like this:

$$E(y_{i3}) = \frac{y_{i2} + y_{i1}}{2} \tag{3}$$

That is, the SDM predicts that the average response is the best predictor of the next response. If change is non-persisting (i.e., if our metaphorical traveler ends up back in her own bed every morning), then taking the average of the last two places we saw her will be our best guess about the location of her home base.

Combined Model

In practice, any data likely contain individuals whose responses reflect both models. Thus, an important question becomes what proportion of observed change in each data set is drawn from each data-generating process. Both of the models above include y_{i2} as a predictor of y_{i3}, but only the SDM model includes y_{i1} as a predictor, and its is treated as equal to y_{i2}. Therefore, the real question is the relative importance of y_{i2} and y_{i1} in predicting y_{i3}. If the two previous estimates are equally predictive, then we are looking at an SDM model. If y_{i2} is a much better predictor, then we have evidence for an active updating process. We therefore combine the previous models into a single model that estimates the relative importance of y_{i2} and y_{i1} in predicting an individual's response at wave 3:

$$E(y_{i3}) = \alpha + \phi\beta y_{i2} + (1 - \phi)\beta y_{i1} \tag{4}$$

Rather than estimate separate parameters for y_{i2} and y_{i1}, this model generates two parameter estimates of interest for each item in our analysis: β, which captures how well any combination of previous waves predicts a person's response at wave 3, and ϕ, the relative proportion of wave 3 explained by wave 2 compared to wave 1. If the Settled Dispositions Model is the preferred data-generating process for an item, then both y_{i2} and y_{i1} should be equally predictive of y_{i3}, and ϕ will equal .5, meaning the best estimate of wave 3 is some transformation of the mean of previous waves, consistent with (3). In the active updating model is preferred, then wave 2 will be more predictive than wave 1, and ϕ will increase to 1 and be consistent with (1).

The Problem of Measurement Error

The discussion above would be completely accurate were it not for measurement error. Measurement error is, by definition, non-persisting change, and as a result, estimates of ϕ will be biased toward .5 in the presence of measurement error. Unfortunately, there is evidence that many of the items explored in our analysis are measured with significant error (Alwin, Hout and Hastings). On the other hand, previous studies of measurement error often conflate measurement error and non-persisting belief change, meaning that while we might have good estimates for the combination of these two processes, we cannot separate them.

Because of this issue measurement error, we cannot compare the relative presence of non-persisting belief changes with persisting changes and therefore the relative prevalence of active updating and settled dispositions models. We can only detect whether there is evidence of persisting belief changes, and therefore whether there is any evidence that active updating present.

Analysis Steps

Our analysis proceeds in three steps. We first evaluate the evidence in favor of the Active Updating model. To do this, we compare for each item the Bayesian Information Criteria (BIC) of a model estimated using with a free estimate of ϕ to a model that constrains $\phi = .5$. If the model with the constraint if preferred, then there is little evidence that respondents engage in an active updating process with respect to that item during the window we study.

Next, for variables that do show at least some evidence of persistent change ($\phi > 0.5$), we ask whether the persistent change is disproportionately concentrated among younger respondents. To test this, we re-estimate our original model and include a dummy term indicating whether respondents fall below an age cutoff. Rather than test a single age cutoff, we again use BIC comparisons to evaluate whether including the dummy variable improves the model fit using a cutoff of every age between 20 and 45. We test a range of cutoffs to ensure robustness of the overall pattern to specific ages. If the cutoff is too restrictive, there might not be enough cases to draw valid inferences. If there are strong differential effects of age in a narrow age range, a large cutoff might dilute the effects of age on active updating.

Finally, with these findings in hand, we consider whether there are any meaningful patterns in the relative distribution of evidence for AUM and SDM across variables.

Item Selection

We seek to test our model on as broad a range of items tapping attitudes, beliefs, and social behaviors as possible. To do so, we estimate the model in equation **??** on data from the three three-wave panel data sets of the General Social Survey, conducted from 2006 through 2014.

We sought measures of attitudes, beliefs, self-assessments, self-perceptions, and behaviors that were asked in three waves. These questions tended to come from the "core" of the GSS, a set of questions asked in each wave. Rotating topical modules asked during the panels were only asked in select waves or were not asked consistently to the same people over time, leaving too few cases with complete data to analyze.

Since our theoretical framework focuses on attitudes, beliefs, and social behaviors, we excluded from our analysis questions that focused on demographic characteristics (marital status, household size, region, gender, race, ethnicity), work activity (employment status, income, hours worked, size of workplace), objective socioeconomic status (years of education and highest degree, home ownership), and evaluations of a respondent by the questioner. We also exclude questions that ask about an individuals' childhood. While testing these questions with our method is possible, they are beyond the scope of our theoretical framework. To organize our presentation of such a large number of variables, we follow Hout and Hastings (XXXX) and group questions into 15 categories based on their subject material. Questions in the same category tend to be asked in the same block during the survey and have the same structure, such as questions about confidence in institutional leaders, questions about public spending, and questions about social life.

We also follow Hout and Hastings (XXXX) in recreating some commonly used scales designed to capture attitudes about gender roles, access to abortion, and social trust. This includes Rossi's six-question scale of support for abortion and a seven-question scale which includes the question asking about abortion under any circumstances (abany). We use Smith's (1997) scale of "misanthropy" by combining questions about how helpful, fair, and trustworthy people are. We use four questions to create a scale of gender role attitudes (Cotter, Hermsen, and Vanneman 2011).

Like Hout and Hastings, we combine civil liberties items into six scales about the freedom of atheists, communists, militarists, racists, and, in the 2010-14 panel, Muslim clergymen. We combine four parallel questions about how frequently individuals socialize to create a "social life" scale. We combine four questions about support for suicide under different circumstances. We also created a scale of support for police use of violence against criminal suspects by averaging five binary questions about the conditions under which individuals support police use of violence.

In total, we test the model on 184 GSS items, including the composite scales.[1]

Results

Our model generates two parameter estimates of interest for each item in our analysis: β, which captures how well any combination of previous waves predicts a person's response at wave 3, and ϕ, the relative proportion of wave 3 explained by wave 2 compared to wave 1. If an item follows an Active Updating Model, both β and ϕ should be close to 1. If responses are generated through a Settled Dispositions Model, then ϕ will be .5.

Evidence for relative importance of AUM and SDM

To evaluate the evidence in favor of the Active Updating model, we compare the Bayesian Information Criteria (BIC) of a model with a free estimate of ϕ to a model that constrains $\phi = .5$. If the model with the constraint if preferred, then there is little evidence that respondents engage in an active updating process with respect to that item.

Figure 1 plots the distribution of ϕ estimates for the 184 items evaluated in this analysis and the probabilities that the model without the constraint fits the data better.

Figure 1 shows that while the majority of items prefer the free parameter, 72 items (about 40 percent of those examined) prefer the constraint, meaning these items show no evidence of active updating during the time frame we examine here. While individuals might shift their response to these items in any particular wave because of measurement error or real ideological change, they will tend to revert to their previously held position over time.

This group of items preferring the constraint includes a broad swath of items across a number of categories. It includes about half of the items about civil liberties, half of the items about confidence in institutional leaders, half of all questions about race and immigration, half the items about gender and family, and half the items about sexuality and abortion. On the other hand, most items about religion, suicide, politics and government, and public spending show evidence of active updating.

Figure 2 plots the ϕ and β estimates for items colored by whether they prefer the constraint or not, with a few items labeled. Items tend to prefer the $\phi = 0.5$ constraint for a combination of two reasons: because wave 1 and wave 2 have equal predictive power or because the measure is so unpredictable (β is so low) that neither wave 1 nor wave 2 has much predictive power. Included in the first group, where waves 1 and 2 are both predictive, are items asking whether a person hunts (labeled in Figure 2), views about the morality of sexual relations among teenagers and extramarital sex, several questions tapping gender roles and gender

[1] To ensure that our estimates of ϕ are not simply artifacts of response scale construction, we estimate the model on coarsened versions of items, generated by collapsing responses to questions with more than three response options into scales of two or three response options. These results are reported in an appendix.

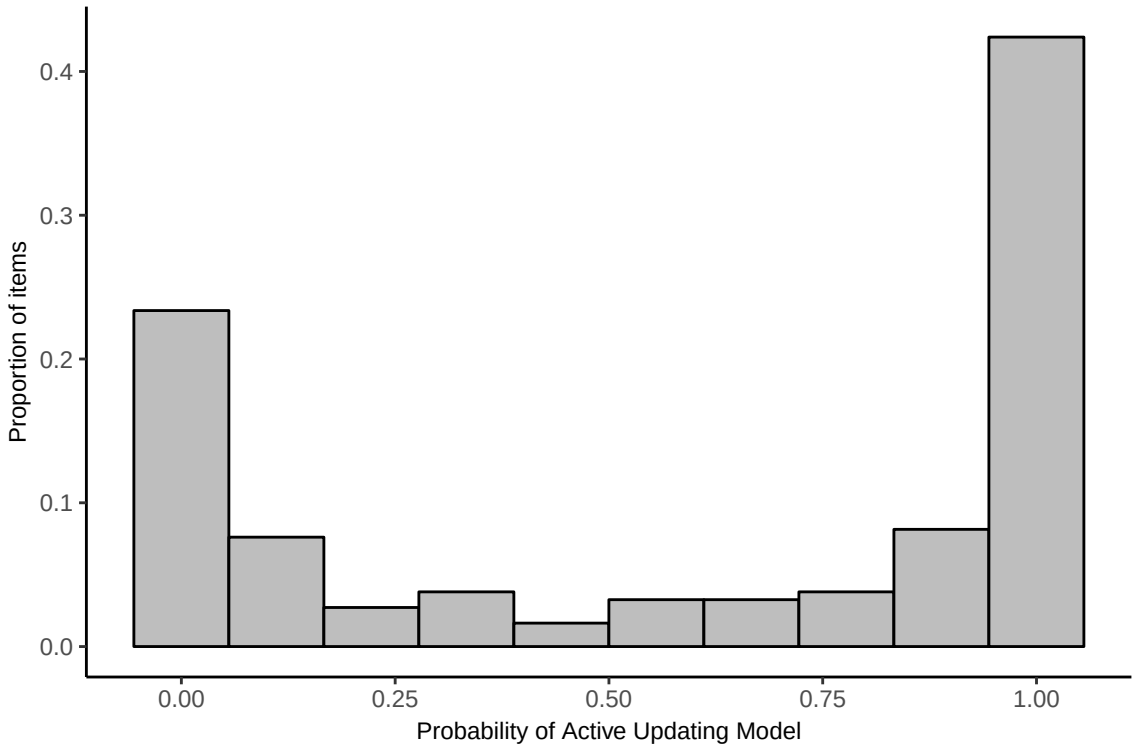

Figure 1: Distribution of probabilities that items favor the Active Updating Model.

discrimination, items capturing stereotypes about race, questions about civil liberties, and most questions and scales about abortion.

The second group of items, those for which neither wave 1 nor wave 2 are particularly predictive and as a result prefer the constraint of $\phi = .5$, comprises a set of questions about racial stereotypes, including whether blacks and whites are intelligent. This is consistent with previous work finding a high level of measurement error for these items (Hout and Hastings). This group also includes items strongly shaped by the Great Recession, such as whether individuals believe they will lose their jobs.

Items preferring the active updating model tend to have ϕ estimates greater than .55, and most have β estimates greater than .6. A small group of variables, including confidence in the leadership of the executive branch of the federal government have low β estimates, meaning that prediction at wave 3 is difficult, but have large ϕ estimates, meaning that wave 2 is still a better predictor.

Age Heterogeneity

The next step in our analysis seeks to explore whether evidence of active updating is disporportionately concentrated in younger respondents.

Of the 112 items that showed evidence for active updating, 23 of these showed differential effects of age for more than 50 percent of cutoff ages. To get a sense of the magnitude of difference between older and younger individuals, Figure 3 plots the estimates of these 23 items for individuals above or equal to or below 30.

The majority of items that show evidence for age concentration show that active updating is more prevalent among younger respondents. These items include views on affirmative action, women in the workforce, and politics, several civil liberties items, general views of whether people can be trusted, and views on whether doctors should let terminal patients die. These items tend to be in categories where a large proportion showed

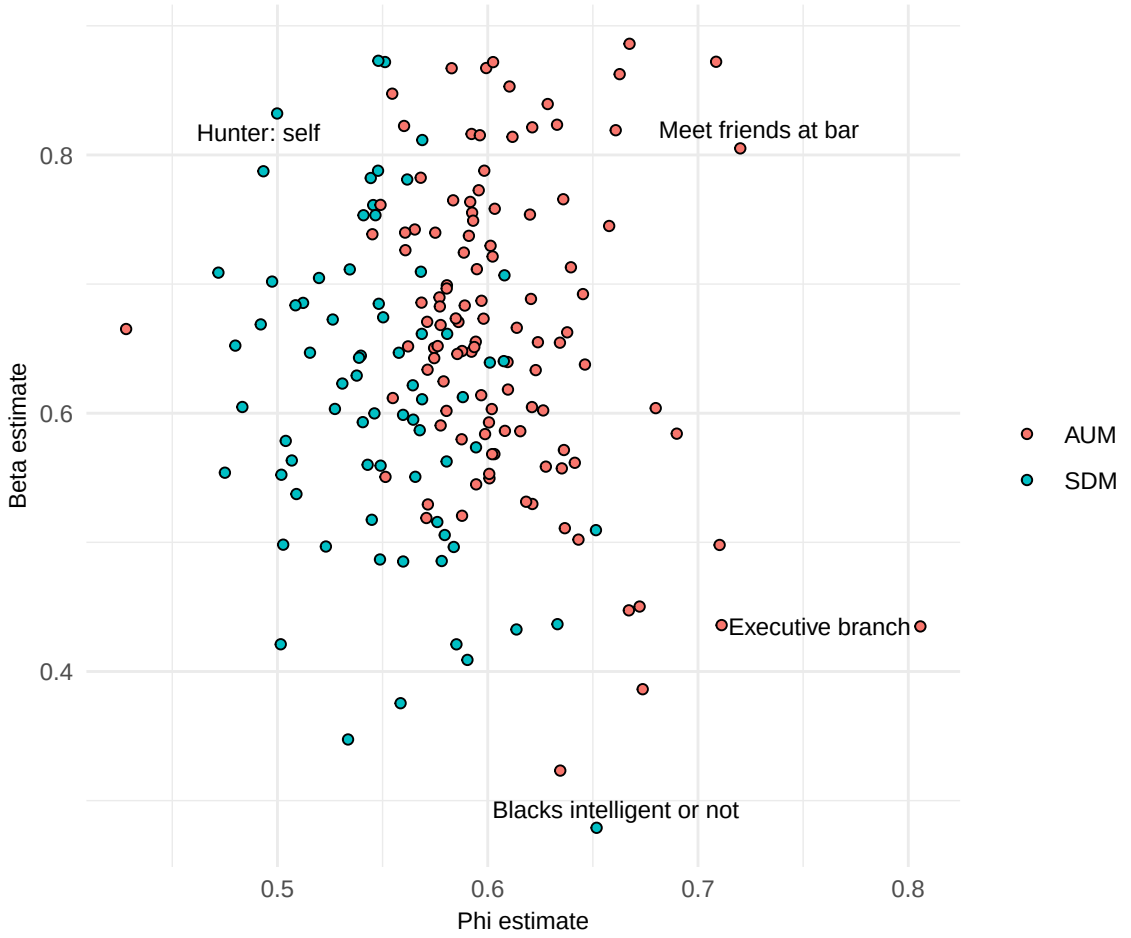

Figure 2: Distribution of beta and phi estimates for GSS items, by preferred model.

no evidence of active updating, which suggests an overall trend of these view being formed earlier in life (perhaps prior to inclusion in the GSS) and remaining relatively stable for the rest of life.

For some items, such as whether individuals can be trusted, political views, whether physicians should allow terminal patients to die, and whether companies should make special efforts to hire and promote women to address pass discrimination, all evidence of active updating disappears for individuals over 30. For other items, such as how important people believe it is for children to be popular and views on how much the government should spend on health care, there is still evidence of active updating in older individuals even though it is substantially less than for younger individuals.

Eight of these items show a negative effect of being below the age cutoff on the ϕ value, meaning that younger people showed less evidence of active updating. These items include how often individuals were active in religious activities, views on suicide in the case of bankruptcy, views on whether aging parents should live with their children, and views on whether police should be allowed to use force against suspects who are verablly abusive toward them.

The remaining 89 items explored in this analysis (just under half of all items) show evidence for active updating but do not show consistent age heterogeneity. However, this does not mean they show strong evidence of an active updating model. We now turn to exploring the extent of evidence for active updating for items in the data set.

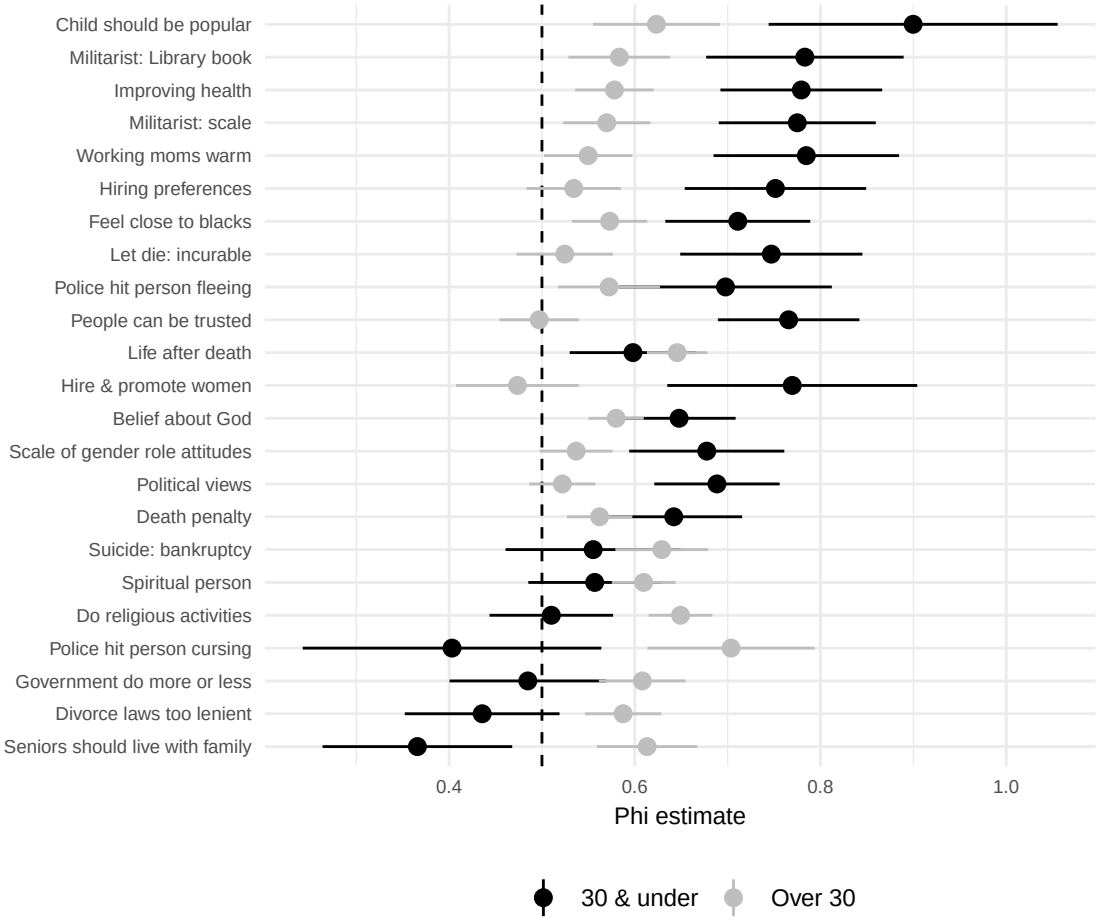

Figure 3: Comparison of phi estimates for individuals over and equal to to less than 30 years old.

Item heterogeneity

Which items show the strongest evidence for active updating? Figure 4 through 7 plot ϕ estimates for all items included in the analysis, grouped by subject material. We constrain items that showed no evidence of active updating to $phi = .5$ in these plots. For ease of interpretation, we omit from these plots one item that showed a statistically significant negative ϕ estimate: whether respondents can imagine situations in which they would approve of a policeman striking an adult male citizen.

The overwhelming takeaway from these figures is that even for those items where the AUM model is preferred, the majority of change in these items is non-persisting, either because there is significant measurement error in individuals' responses or because individuals make non-persisting changes at higher rates than they make persisting change. The majority of items that show evidence of active updating have ϕ estimates less than .6, with only five items exceeding .7. Items that show the highest rate of active updating include items with clearly identified external referents that change (e.g., confidence in the executive branch of the federal government), items that were strongly influenced by the Great Recession of 2008 (a person's assessment of whether their financial situation has changed and ability to get a new job), and facts such as whether someone owns a gun, which are easily recalled and tend to have low levels of measurement error.

Because change can take on a variety of forms, and because the estimate of ϕ varies depending on the direction and relative variance of real change and measurement error, these ϕ estimates do not translate clearly into an interpretable metric. For reference, however, in simulated models where persisting and non-persisting change

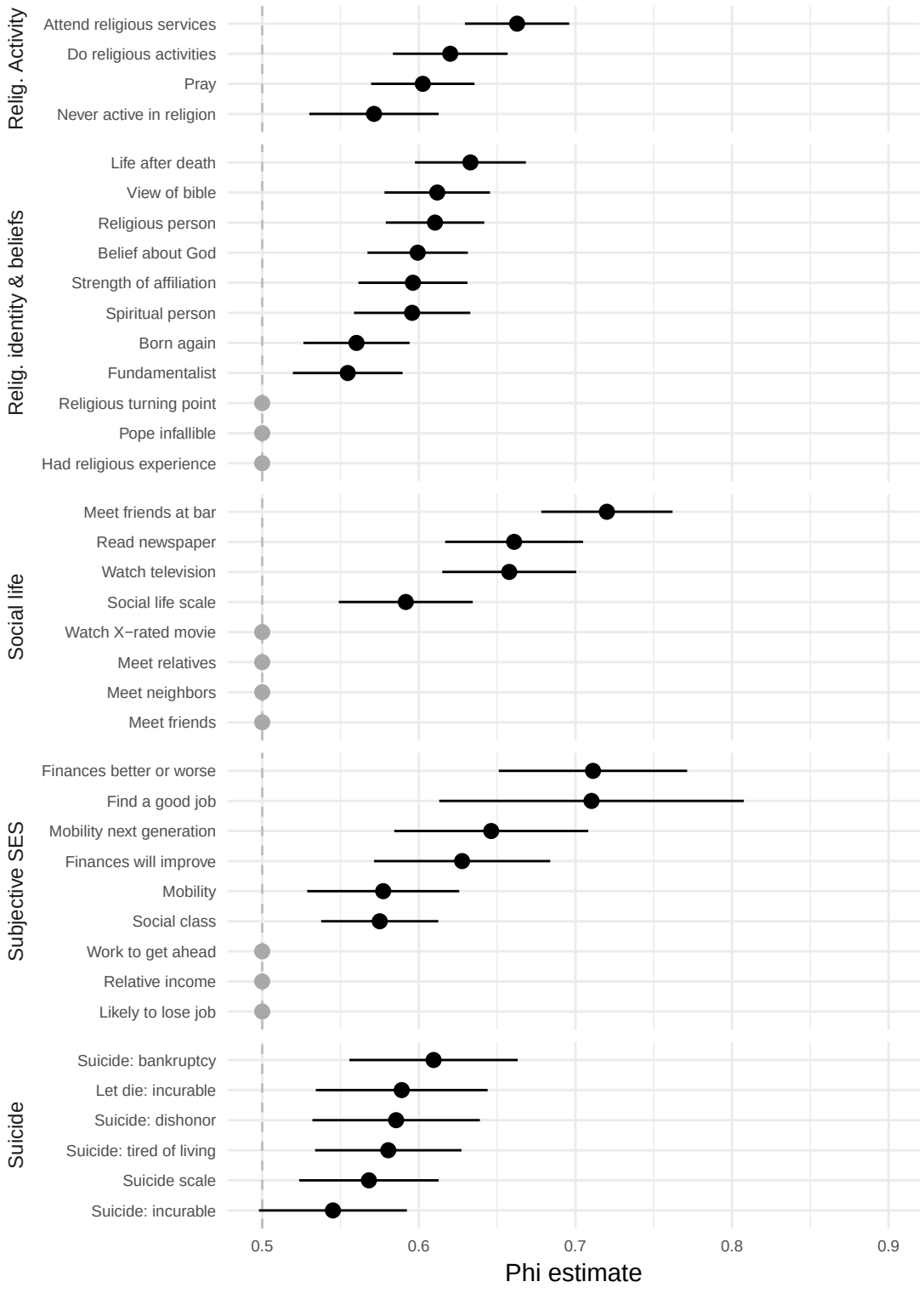

Figure 4: Estimated conditional proportion of persistent change (phi) for full panel data set and unconditional proportion of persistent change estimate (phi*delta) for items about religious activity and beliefs, social life, subjective SES, and suicide.

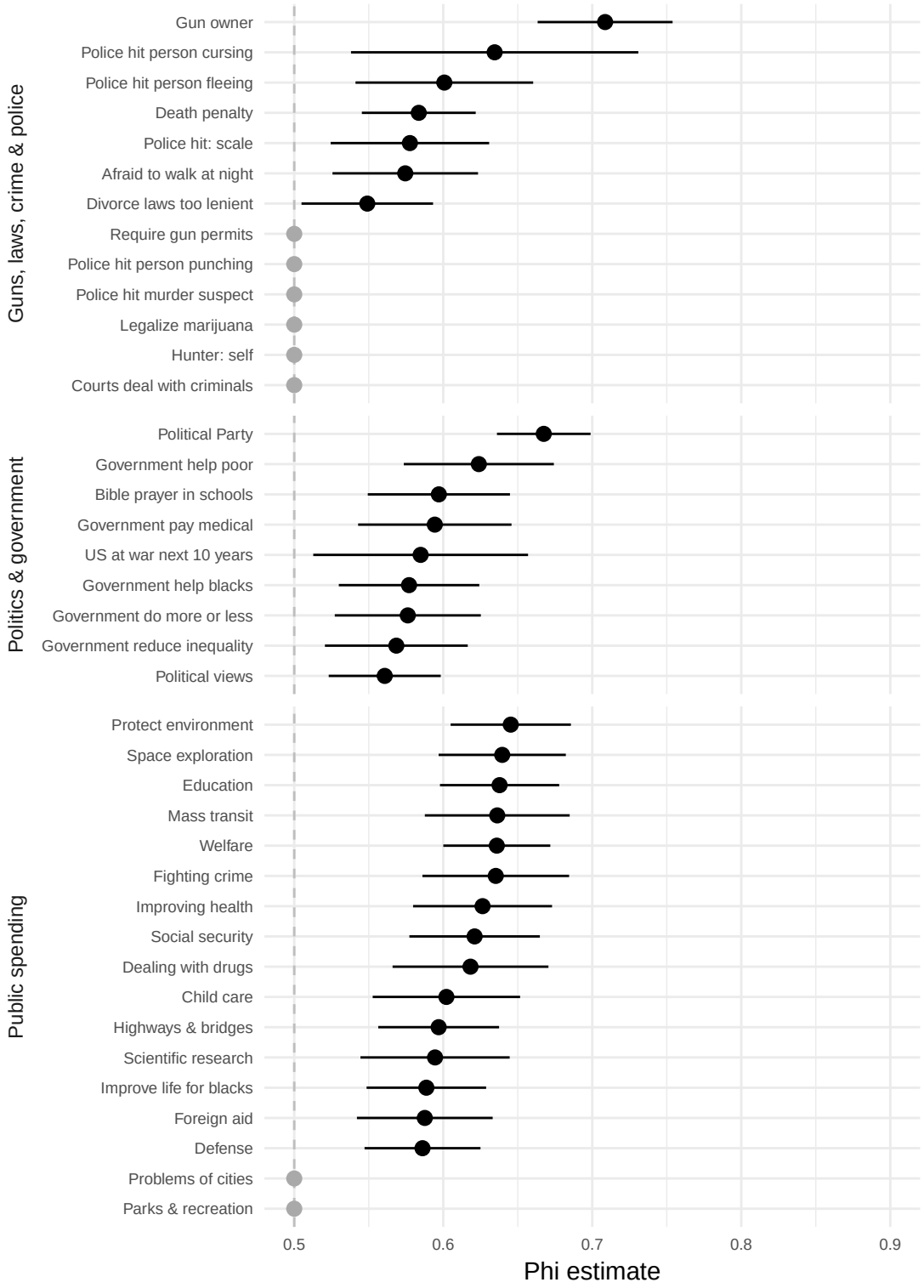

Figure 5: Estimated conditional proportion of persistent change (phi) for full panel data set and unconditional proportion of persistent change estimate for items about guns, law, crime and policing; politics and government; and public spending.

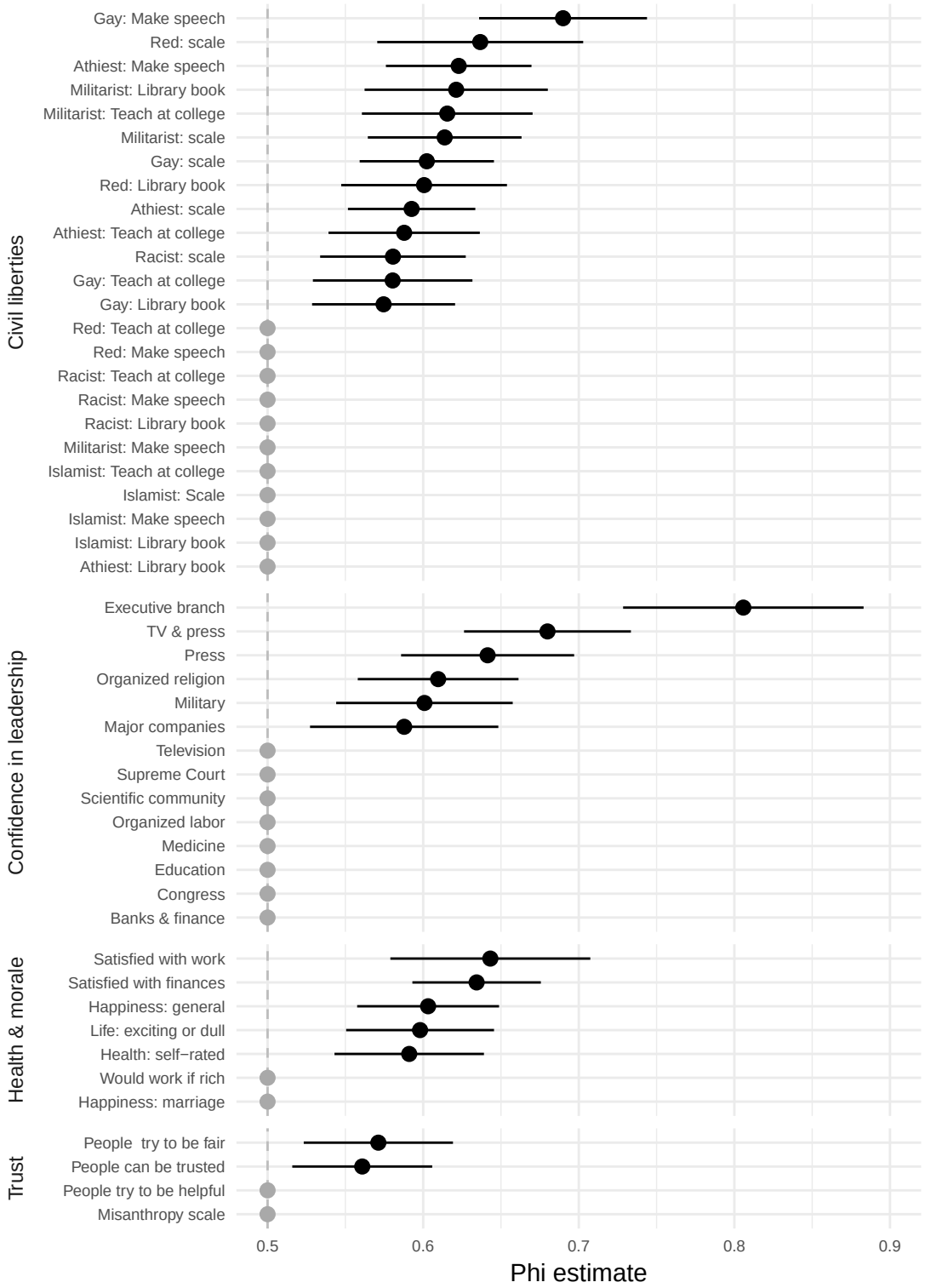

Figure 6: Estimated conditional proportion of persistent change (phi) for full panel data set and unconditional proportion of persistent change estimatefor items about civil liberties, confidence in leadership, health, morale, and social trust.

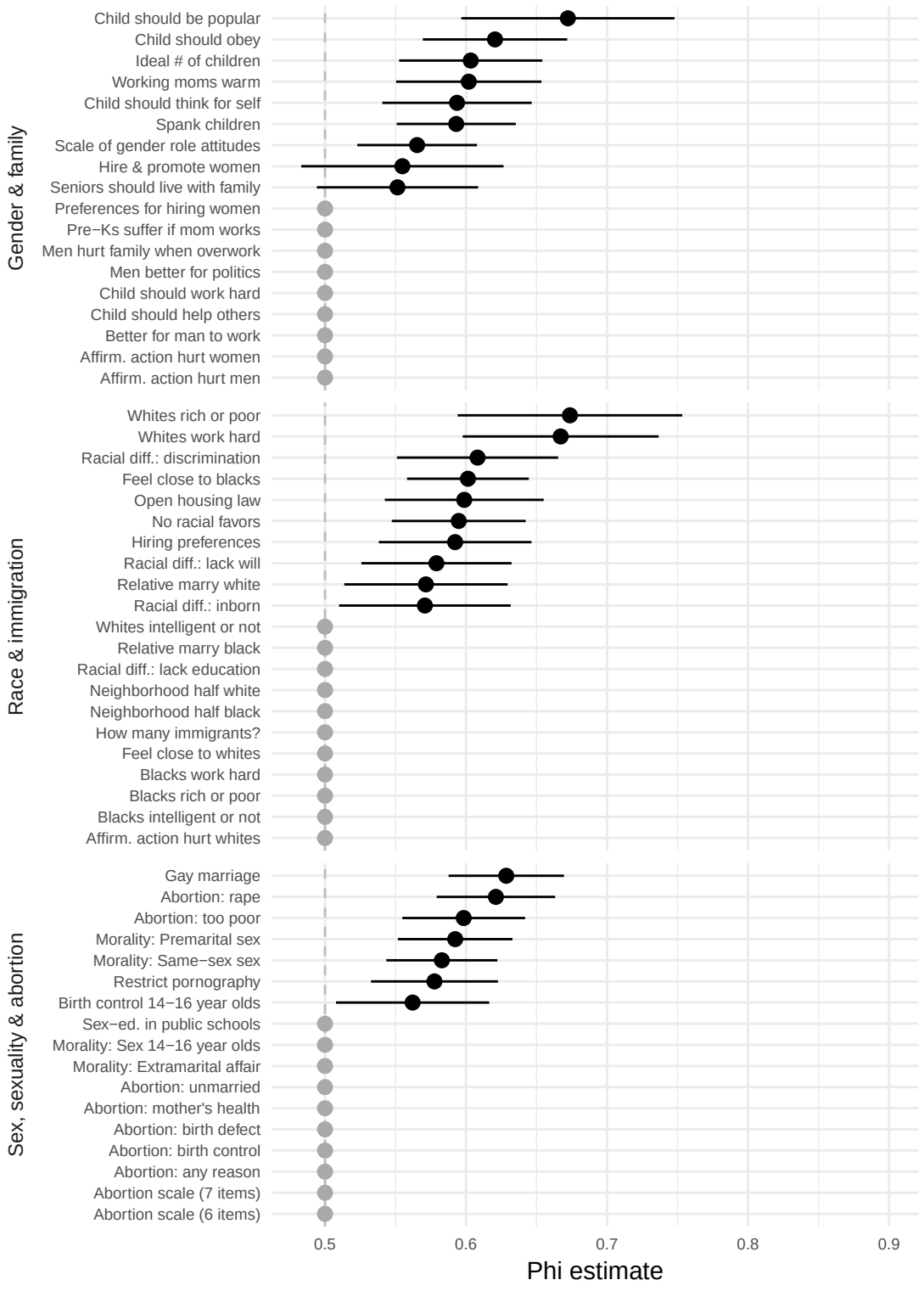

Figure 7: Estimated conditional proportion of persistent change (phi) for full panel data set and unconditional proportion of persistent change estimate for items about race, gender, sex, sexuality, and abortion.

are drawn from the same distribution with no measurement error, a ϕ of .6 equates to about twice as many people making non-persisting changes as persisting changes.

It's impossible to talk about all results in these tables, but we highlight some notable results and trends. Starting with Figure 4, religious activity is the only category where all items show evidence of an active updating process, and these items appear to have stronger evidence than other kinds of questions, with $\phi > .6$ for three of the four items. Some items that we might assume would have high evidence of persistence do not. Whether a person had a religious turning point or was "born again" show no evidence of persistence, a surprising finding given that these changes should be non-reversible.

Almost all items dealing with politics, government, and public spending show evidence for an active updating model, with ϕ estimates close to .6. Within items dealing with politics, which political party an individual identifies with shows more evidence than other questions for active updating, while general political ideology and views on the government's role show less evidence.

Within the civil liberties category, persisting change does not appear to be concentrated among either a type of freedom (library, speech, college) or a subject of that freedom (atheist, militarist), as types and subjects see both persisting and non-persisting change. The notable exception here is items about freedom for homosexuals, all of which show evidence of persisting change. Similarly, all but one of the scales show evidence of persisting change.

Within the "confidence in leadership" category, low-visibility institutions with harder-to-identify leadership, such as the scientific community, education, organized labor, and medicine, tend to have no evidence for active updating, while high-visibility institutional leaders such as organized religion, the financial system, and the executive branch of the federal government show evidence of active updating.

As noted previously, gender, race, and sexuality topics in Figure 7 tend to have a higher proporton of "settled" items compared to other categories. Even for those items that do show evidence of persistence, the ϕ estimates tend to be smaller (less than 0.6), with the exception of items such as support for same-sex marriage and support for abrtion in the case of rape. Within items about sex and sexuality, all but two items on abortion appear to be settled.

As noted previously, low measurement error might be a reasonable assumption for some items. Previous studies using different approaches to measuring the reliability of survey reports suggest that some items captured in our study, such as whether a person owns a gun, are measured with a high degree of reliability ($>$ .9) (Hout and Hastings). For other items, such as confidence in the leadership of major compaines ($\phi = .58$), reliability might be as low as .5.

While we cannot control for measurment error in our analysis, we can take steps to mitigate its impact. Appendix A presents results comparing items with more than three scale points to coarsened versions of these question with either two or three scale points. If scales had symmetrical scales with no clear midpoint (e.g., strongly agree/agree/disagree/strongly disagree), we coarsened those scales to two points (agree/disagree). If scales included a clear midpoint, we included that and coarsened responses on either side. For items on large scales such as hours of tv watched, we split responses into greater than or less than or equal to the median.

Generally speaking, most items with no evidence of active updating continued to have no evidence of active updating when coarsened. For items that showed evidence for active updating, the coarsened estiamtes tended to be very similar to the uncoarsened estimates, suggesting that persisting and non-persisting changes happen about as often between ends of these scales as they do within ends of these scales.

There are a couple notable departures from this general pattern. Several political questions – general political views, views on whether the government should reduce inequality, and views on whether the government should help blacks – show decreased evidence of persisting change when coarsened. This suggests that changes around the midpoint of the scale tend to be measurement error, a finding that's consistent with previous work suggesting that individuals without settled political views tend to choose points in the middle of the response scale.