DOI: 10.1111/pops.70056

SPECIAL ISSUE ARTICLE

Political Psychology

The promises and pitfalls of using panel data to understand individual belief change

Turgut Keskintürk | **Pablo Bello** | **Stephen Vaisey**

Department of Sociology, Duke University, Durham, North Carolina, USA

Correspondence
Turgut Keskintürk, Department of Sociology, Duke University, 276 Reuben-Cooke Building, Durham, NC 27708, USA.
Email: turgut.keskinturk@duke.edu

Abstract

We investigate whether studies on political belief change can identify change trajectories at the individual level. Using simulations and case studies, we propose a grid-search framework that allows researchers to evaluate the extent to which their target estimates generalize to their study population. We use simulated datasets to estimate plausible values for how many people changed, how much they changed, and who changed, based on observed response trajectories. Our results suggest that researchers should think carefully about the conditions under which they may make claims about belief change at the individual level. To guide substantive theory-building, we propose a concise diagnostic routine researchers can use to translate their claims into a set of plausible alternatives and evaluate potential generative processes. We provide an R package to help researchers implement this procedure in their own work.

KEYWORDS

belief change, panel data, survey methodology

Studies on political change at the individual level—change in one's beliefs and preferences—face a challenging problem: while our methods typically provide estimates for change and stability at the population level, our goal is often to understand these processes at the individual level.

This *group-to-person generalizability problem*, the claim that group-level findings may not generalize to each person (McManus et al., 2023), has important implications for what we can and cannot say about political change across the life course. While previous research has quantified change with various methods—ranging from assessing the time order of longitudinal observations (Kiley & Vaisey, 2020; Vaisey & Kiley, 2021) to multilevel decomposition of over-time variance (Lersch, 2023)—these studies have been unable to disentangle whether they find a small amount of change in a large number of people, a large amount of change in a small number of people, or even some mix of positive and negative trajectories. In other words, while

social science researchers have become sophisticated in evaluating the average levels of change over time in a population, they have mostly overlooked the fact that these average changes can be produced by different processes.

This issue arises in many ways in empirical research, particularly when we want to theorize the mechanisms of political belief change. For example, the US General Social Survey (GSS) shows that roughly 35% of the American adults in 2006 believed that "marijuana should be legalized." In less than 10 years, this support had increased to nearly 60%. As seen in Figure 1, this change occurred both *cross-sectionally*, where average support increased with time, and *longitudinally*, where at least some segments of the American population changed their minds on this policy over these periods.

However, these two processes, cross-sectional change and longitudinal change, are connected to competing theories of political change, and group-level estimates may not properly adjudicate the underlying data generation processes. Staying with the example of marijuana legalization, the observed cross-sectional change in beliefs may have occurred via two processes: people may have changed their minds following a changing *zeitgeist*, or older generations who disapproved of this policy might have been replaced by new generations approving it (Ryder, 1965; e.g., Small, 2002). Similarly, the longitudinal change could result from various processes: large segments of the population may have shifted their opinions positively in small amounts, a small segment of the population may have made large changes, or a mix of changes in different directions may have altered the balance.

While cross-sectional processes have received strong scholarly attention in the literature (Bartels & Jackman, 2014; Ochoa & Vaisey, 2024; Vaisey & Lizardo, 2016), questions of the second sort have remained largely unexplored. However, given that theories of political ideology involve theoretical expectations about individual-level outcomes, the fact that we lack information about whether our models capture person-level empirical processes is detrimental to understanding adult political belief change. Our article aims to address this gap by examining whether we can identify (a) *how many people changed*, (b) *how much they changed*, and (c) *who changed* using empirical data.

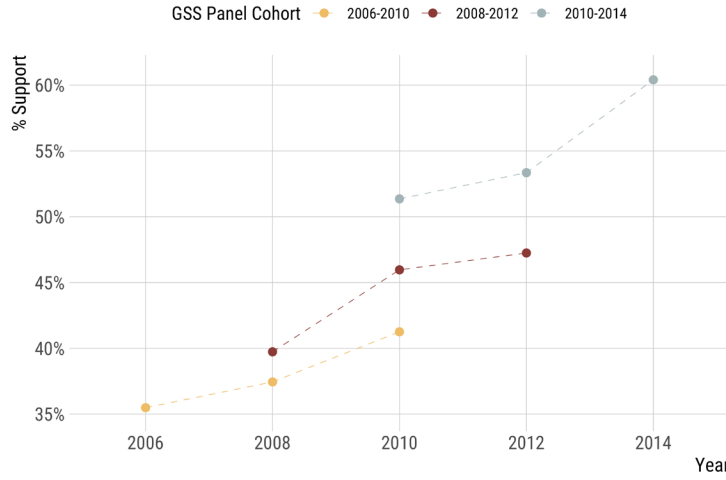

FIGURE 1 The percent support for marijuana legalization. The figure depicts the percent of American adults saying "marijuana should be made legal" in the General Social Survey 2006, 2008, and 2010 panel studies. We included respondents who participated in all three waves of each panel study and weighted the results using GSS post-stratification weights. Note that panel attrition caused compositional changes in the estimates, and compared to the GSS time-series data, the estimates of support are marginally inflated.

Using simulations and empirical case studies of political change, we show that it is indeed possible to use panel data to provide plausible estimates of which generative models are more likely to have produced particular empirical observations, even if identifying exactly which model produced the data is implausible. We also find, however, that a reliable identification of who changed occurs only under very narrow conditions. Hence, the problem of modeling individual change is very much like the age–period–cohort identification problem that arises in repeated cross-sectional data (Fosse & Winship, 2019). In both cases, the problem cannot be "solved" by any particular methodological trick. Because of this, our objective is *not* to provide a methodological panacea to solve the problem, but rather to provide a concrete approach that can help us understand the issue more clearly and narrow the range of possible solutions explaining change trajectories in a given dataset.

This approach builds on the "Approximate Bayesian Computation" (ABC) literature (Beaumont, 2010) and provides a simple grid-search procedure to help researchers evaluate the extent to which their target estimates might be generalized to a study population. Using this procedure, we show that it is possible to approximate the set of underlying data generation processes for *rate of change*, *strength of change*, and *direction of change* in a panel dataset. We provide an R package, "gridsearch," to help researchers implement our ABC diagnostic in their own work.

We emphasize the existence of striking heterogeneity in the set of plausible underlying generative processes, even for a seemingly straightforward case like the marijuana legalization in the GSS. That is, we show that opposite claims—for example, that large swaths of the population increased their support for the policy or that a small group of people completely switched their positions—may very well account for the same data. This finding is particularly relevant in the longitudinal panel context, which, for most applications, represents the gold standard in evaluating change in observational research. For this reason, we argue that when the meaning of "political change" is not clearly specified, our current empirical strategies could potentially mislead rather than inform.

We recommend our grid-search routine as an initial diagnostic protocol to guide theory-building and substantive interpretation. Concretely, we propose that researchers should (1) identify which generative processes may have produced the distribution of their observed outcomes; (2) use the resulting equivalence class of plausible processes to specify a small set of theoretically defensible generative models; (3) guide the analytic choices and substantive interpretations that follow. Accordingly, our goal is to provide a transparent checklist that helps researchers recognize when summary statistics may obscure heterogeneous individual-level dynamics in panel data.

We begin with a general overview of the "latent variable model" of beliefs, the problems of reliability and resolution in survey measurement, and the compositional sources of aggregate change. We then introduce our grid-search approach with an illustrating exercise. We provide empirical examples and explore the implications of our approach. Next, we analyze whether our simulation approach helps us in classification tasks, particularly in evaluating *who* changed. We conclude by discussing the promises and pitfalls of panel data for understanding individual political change.

MEASURING CHANGE IN POLITICAL BELIEFS AND PREFERENCES

The latent variable approach to survey response is based on the assumption that there is a random variable Y, which captures a person i's "true" position in a continuous latent

distribution (Alwin, 2007). This random variable might stand in for any disposition, including one's political beliefs and preferences, which allows us to define one's political position as their *true score* Y_i.[1]

The latent variable approach acknowledges a fundamental problem in survey research: any measure of one's true score Y_i is almost always realized with a certain degree of measurement error. Call this realization the *observed score*, y_i. We often assume that y_i represents the expected value of hypothetical draws from Y_i, with an ϵ term reflecting the measurement error in this process—that is, $y_i = Y_i + \epsilon_i$. This error might result from various factors that introduce biases in how we extract information from individuals, given that problems like question construction and interviewer effects might influence how individuals reveal their true preferences (Alwin & Krosnick, 1991). The inherent randomness in the measurement process implies that each draw from Y_i might at least slightly differ from the true score, and this error term is difficult to circumvent with a one-item measure, which is often the case in large public opinion surveys. Therefore, there is often a degree of systematic gap between a true belief or preference, Y_i, and one's survey response, y_i.

In the panel context, where we observe the same people multiple times, this measurement problem has substantive implications: if true scores are imperfectly realized and we only have access to a person's observed survey responses, how can we know if they *actually* changed between time t and $t + 1$, or whether the observed change is merely due to measurement error? For instance, when we claim that support for abortion "rose an average of 28 points by mid-1970s" (Hout et al., 2022, p. 2), or preferences for redistribution changed among 50% of the GSS respondents who became unemployed between 2006 and 2008 (Owens & Pedulla, 2014), how can we be sure that these changes represent true changes in people's latent positions, rather than a simple error?

Scholars in public opinion research have attempted to address this issue in multiple ways, using structural equation models that explicitly quantify the measurement error component (Achen, 1975), strategies that integrate longitudinal surveys to exploit the information in the ordering of responses (Kiley & Vaisey, 2020), and various multilevel mixed effects models that shrink responses based on population estimates to derive reliable trends (Hout & Hastings, 2016; Lersch, 2023). To some extent, all these strategies have been useful for offsetting the problems associated with measurement error.[2] That said, while these empirical strategies have been largely successful in accounting for the amount of change *in the population*, they were rather silent about changes *at the level of individuals*.

The compositional sources of aggregate change

Suppose we have a panel data, where individuals $i = 1, \ldots, N$ are observed over two time periods t_0 and t_1. Let y_{it} represent an observed survey response. To see how much observed change there is at the individual level, we can simply define a change quantity, $\delta y_{i,t_0 \to t_1} = y_{it_1} - y_{it_0}$, the average of which, $\overline{\delta}_y$, represents the mean change in the dataset from t_0 to t_1.

[1]Naturally, the assumption that *there is indeed a true score* might miss cases where individuals hold contradictory ideas (Swidler, 1986) and are ambivalent about their true positions (Zaller, 1992; Zaller & Feldman, 1992). In this article, we follow the latent variable model and assume that there *is* a true position for each individual; hence, bracketing the cases where this ambivalence needs to be incorporated into the survey response measurement model.

[2]However, note that the use of repeated observations in the panel context comes with additional assumptions: it is rather straightforward to assume that multiple measures at time t can be indicative of an underlying construct—for example, we can average them to have an index—but the use of different time periods needs a *stability* assumption, that is, the reliability estimates may only be computed if one assumes that people do not really change between waves, which might confuse true change with measurement error (see, e.g., Hout & Hastings, 2016).

This quantity is a natural candidate for describing aggregate belief change, but, under the hood, it summarizes three compositional processes. Let us define these processes intuitively. The first component is R, or the *rate of change*. This represents the proportion of individuals in our dataset with $\delta y_{i,t_0 \to t_1} \neq 0$, that is, the proportion who changed in this period. The second component is S, the *strength of change*. This represents the average magnitude of change among those who changed. The third component is D, the *direction of change*. This represents the net directionality of change, that is, the difference in proportions of positive and negative changers in our observations. If we combine these three components, we can represent our aggregate change measure, $\overline{\delta}_y$, simply as

$$\overline{\delta}_y = R \times S \times D$$

This simple decomposition should let us appreciate that the average change score may be generated via multiple data generation processes, each having strongly different theoretical implications.

For instance, imagine we observe that society has become more favorable to redistributive policies over time. This change might have occurred via at least three different generative processes. It could be the case that most people in the population are slowly becoming more pro-redistribution. It is also plausible that a small group of those who previously held negative views on distributive policies had swiftly changed sides, while the preferences of most others remained stable. One other potential is a mix of processes: while some people may be moving in the positive direction, others may move negatively, and the net score is simply the difference in their trajectories. All of these processes can produce, at the population level, the same level of change ($\overline{\delta}_y$) even though it is clear that they are substantively different in a way that points to competing theories of cultural change.

The same three processes could also explain the increasingly favorable attitudes toward marijuana legalization that we discussed above. Similar processes might be raised for confidence in legislative groups and financial institutions, beliefs about the economy, and even gender attitudes. In this sense, the capability of our models to distinguish between alternative processes is more limited than is often recognized. Although similar identification problems are widely recognized, like the age-period-cohort (APC) problem, our decomposition of change into rate, strength, and direction and the limitations it implies for observational studies are not generally acknowledged.

But if $\overline{\delta}_y$ obscures these processes, why not directly look at individual responses? This would only be a solution to the problem if we were to believe that each and every single survey response y_i perfectly represents the true position of the individual Y_i. We contend that issues of measurement in survey responses do not allow us to do that. The measurement of individual beliefs often fails in at least two ways: (1) our observed measures might have low reliability, such that a hypothetical set of repeated measurements of the same person may significantly vary around one's true score, and (2) the question response categories might be imperfect: the resolution of responses may fail to distinguish between consequential differences in true scores. Therefore, the average change captured in $\overline{\delta}_y$ represents a biased or imperfect estimate of "true" change to begin with.

We can now clearly state our research question: given a panel study, how can we determine the underlying generative process—that is, how many people changed, how much their latent positions changed, and who changed—given a level of latent score reliability, response resolution, and a longitudinal survey design? In other words, in multiple empirical scenarios, how well can we identify the *true* process that generated our data given a set of conditions?

We pose this question in a longitudinal panel data context (Allison, 2009; Vaisey & Miles, 2017), with varying individual trajectories across time periods (Rüttenauer & Ludwig, 2020). This allows us to shift the focus from *aggregate change* to *individual change*, given that our theoretical interest is to examine how people change their political beliefs across the life course.

A grid-search procedure to adjudicate data generation processes

We build on the Approximate Bayesian Computation (ABC) literature (Beaumont, 2010; Sisson et al., 2018) to propose a grid-search procedure that can answer our research question. ABC is a simulation-based approach to parameter estimation that follows Bayesian logic. Simulation-based approaches have become popular in disciplines like biology and physics. They are used in situations where conventional statistical models—based on a likelihood function—are hard to specify. ABC replaces the traditional likelihood function for a simulation model. The simulation model is used to generate data that is then systematically contrasted to the observed data. The "Bayesian" in Approximate Bayesian Computation implies that each parameter of the model needs a prior distribution and the output of the model is a set of posterior distributions.

We use a rejection sampling algorithm to carry out this procedure, which consists of the following steps (see Sisson et al., 2018 for an introduction to ABC and alternative approaches to estimation):

1. Create a summary function for the observed data. In the case of a binary individual belief, we summarize the observed data with the distribution of individual-level change scores. We obtain this distribution by fitting a logistic regression for each individual, where the outcome is whether or not they hold the belief or preference of interest. More precisely, we estimate:

$$log\left(\frac{\mathbb{P}(y=1)}{1-\mathbb{P}(y=1)}\right) = \beta_0 + \beta_1 t, t \in [0,1]$$

   where t is a normalized time variable. Based on this model, we predict the difference between $t = 1$ and $t = 0$ of holding a belief ($y = 1$), which we refer to as the "change score." Because each individual has their own change score, these form a distribution over the sample of the study. This distribution is the target that we want to approximate with our simulations.
2. Simulate data based on a specific generative model. In this step, we define a simulation model that can accommodate different substantive processes of change, or DGPs, described above. This model has four parameters: rate of change, strength of change, direction of change, and reliability of response measurement. Based on a particular combination of these parameters, we simulate a dataset and calculate the same summary function as in step (1), but for the simulated data in this case. Formally, our model is defined as:

$$y_{it} = Y_i + \delta\tau_i D_{it} + \epsilon_{it} \text{ with } \epsilon_{it} \sim \mathcal{N}(0,1)$$

   where we have a random variable Y_i for true scores, with varying realizations y_{it}, representing a person i's beliefs at time t; an indicator for change, D_{it}, operationalized as a non-reversible trigger if there is a respondent-level change in true scores; an effect size, δ, and a respondent-level direction multiplier τ_i that indexes whether the change is positive

or negative.[3] The parameter δ represents the strength of change, D_{it} relates to the rate of change, and τ_i to the direction of change. Therefore, particular combinations of values for these parameters characterize distinct processes of cultural change. For instance, a low value of δ (indicating small changes in beliefs), combined with a high proportion of individuals for which $D_{it} = 1$ (most people are changing their beliefs) and $\tau_i = 1$ (they change positively), corresponds to a process of a slow but widespread increase in support for a belief.[4]

3. Compute the distance between the summary of the observed data and the simulated one. We use the Kolmogorov-Smirnov statistic to measure the degree of overlap between the real and simulated distributions of change scores to calculate this distance measure.
4. Replicate steps 2 and 3 N times, each one drawing new samples from a proposal distribution for each parameter (i.e., rate of change, strength of change, direction of change, and reliability). The proposal distribution determines the range of values for each parameter.
5. Take the samples that resulted in the smallest distance as plausible DGPs for our observed data. Smaller distances indicate a closer resemblance of the simulated to the observed data, indicating that the proposed DGP could have generated our real-world data.

We refer to this process as *grid search*. The algorithm systematically combs through combinations of parameters looking for those more likely to have generated the data. These combinations of parameters, or DGPs, can be represented as a multidimensional grid in which each parameter represents a new dimension, and each combination of parameters is a point in the grid; hence, its name. This procedure is powerful: it allows us to make educated inferences as to what kinds of processes may or may not be at play in our data while being systematic and principled.

Approximating observed data via simulations

Let us flesh out our approach by walking through an illustrative example of how we approximate the observed empirical responses. Suppose that we *know* the underlying DGP of a dataset of interest. In this dataset, called DGP 1, 500 people are observed over a period of 12 waves. We know that 50% of these respondents make certain changes to their latent positions, and these changes are all in the positive direction with a strength of 1 SD. However, we also know that the item reliability is 80%, which means that, across time, we sometimes miss respondents' true scores. The top-left panel of Figure 2 shows the distribution of change scores from this dataset.

Assume that this dataset is our *observed* dataset. We will now see whether an alternative dataset, called DGP 2, can approximate its observed parameters. From DGP 1, we know that there are 500 individuals observed across 12 waves. We also know the marginal distribution of our response variable. We can treat these parameter values as fixed for our new simulation. However, we do not know several other parameters: we do not know the *real* rate of change, nor do we know the *real* strength, direction, or reliability. Let us assume, for this specific iteration, a rate of 50% (correct), a strength of 1 SD (correct), a reliability score of 80% (correct), and

[3]Note that in this model, actors can change at any wave of the study, making D_{it} variable over t.

[4]We define additional parameters that we fix in the case studies presented below: N refers to the number of actors in the simulation, T refers to the number of waves we have in the panel study, *balance* refers to the observed marginals in the real datasets, allowing us to "cut" the continuous distribution to 0s and 1s, and *resolution* refers to the number of categories to simplify the latent variable, which is restricted to the binary case in this article.

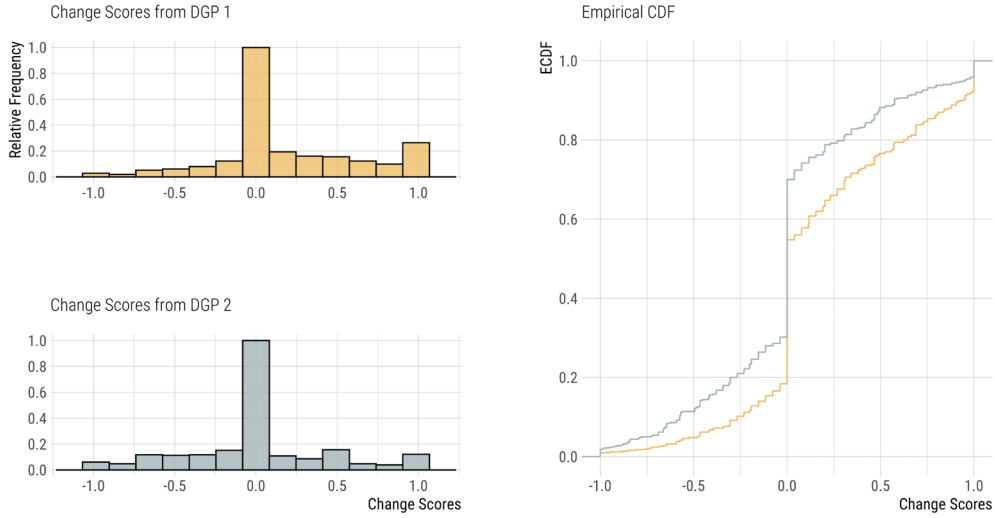

FIGURE 2 Calculating distance for observed and simulated trajectories. The figure presents an illustrative example of two known DGPs: (1) the top-left panel shows the distribution of change scores from DGP 1, (2) the bottom-left panel shows the distribution of change scores from DGP 2, and (3) the right panel shows the empirical cumulative distribution functions from each of these distributions.

an equally split direction, half changing positively and half changing negatively (incorrect). The bottom-left panel in Figure 2 shows the distribution of change scores from this particular incorrect DGP.

How adequately does DGP 2 represent DGP 1? In the right panel of Figure 2, we show the empirical cumulative distribution functions (ECDFs) for our two distributions, which indicate that DGP 2 does a very poor job of reproducing DGP 1, with a Kolmogorov–Smirnov distance of .16. This means that we should be able to improve our predictions significantly if we start tweaking some parameters for DGP 2. Unfortunately, we of course do not know which parameters to tweak; the set of true change parameters is the exact thing we want to derive using observed trajectories. This leads us to consider a wide variety of parameter values that minimize the Kolmogorov–Smirnov distance scores between the observed trajectories and the said simulated trajectories. This is the exact place to use our grid-search procedure to explore a range of DGPs across the parameter space.

EMPIRICAL CASE STUDIES

In the following sections, we apply this procedure in different contexts, analyzing the *group-to-person generalizability* problem across real-world questions on political belief change.

In all case studies and simulation analyses, we restrict our attention to the stability and change of latent variables with binary outcomes to simplify our analyses.[5] The replication files for the article are stored in https://osf.io/5x48d/. The R package for the use of this approach, "gridsearch," can be accessed via the replication repo or https://tkeskinturk.github.io/gridsearch/.

[5]Since we use individual-level models, our approach is generalizable to responses with more than 2 categories. We restrict ourselves to the binary case to keep the mapping of latent scores to observed scores as smooth as possible, so that we do not have to impose additional assumptions that would detract us from the main point of our article.

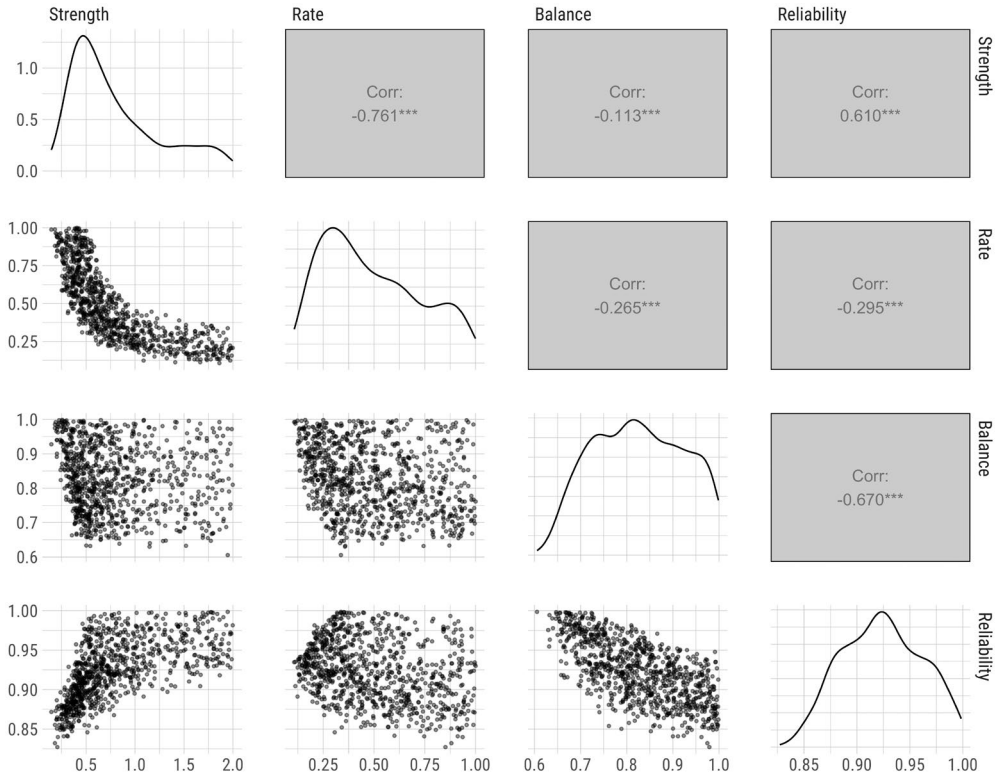

FIGURE 3 Pair plots for marijuana legalization. The figure depicts the distribution of, and bivariate relationships among, grid-search parameter values from the best-fitting DGPs for the case of marijuana legalization (KS distance < .005).

Exploring the case of marijuana legalization

Let us now return to the marijuana legalization case in the GSS and apply our grid-search procedure to see whether we can detect a set of plausible DGPs for our empirical data.

We first fit 1,000,000 simulations by drawing samples from four different parameters: *rate of change*, fixed between 0 and 1; *strength of change*, fixed between 0 SD and 2 SD; *directional balance of change*, fixed between 0 and 1, representing "all negative" change at 0 and "all positive" change at 1; and a *reliability score*, fixed between 60% and 100%. We use uniform priors for each parameter. We then calculate the Kolmogorov–Smirnov (KS) distance for each draw and keep KS distance values less than .005, ending up with 911 potential generative models that might explain our observed data.

Figure 3 shows the distribution of, and bivariate relationships among, these grid-search parameters. Each diagonal panel presents the posterior for a single parameter. The lower-triangular scatter plots compare every pair of parameters, where darker clouds indicate combinations that yield a very small KS distance (i.e., a better fit). The upper-triangular panels provide their correlations. The figure suggests that the posterior distributions of strength and rate parameters are highly variable—and strongly correlated—while the posterior distributions for balance and reliability are much more concentrated. Looking at the distributions, the interquartile ranges for strength and rate are .57 and .37, respectively, while the values are only .16 for balance and .06 for reliability.

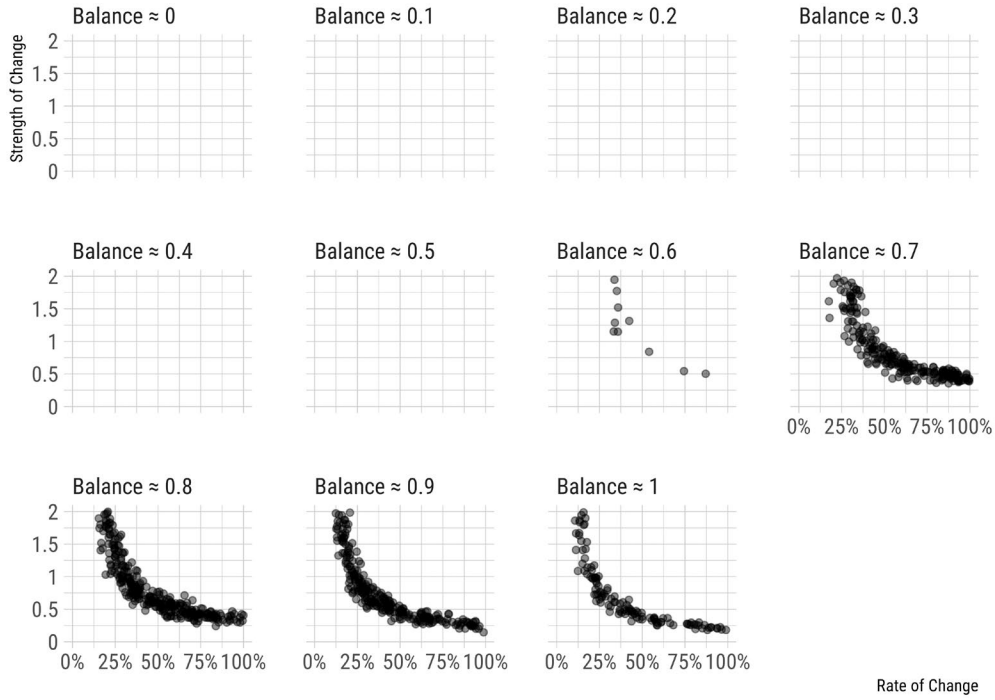

FIGURE 4 Best-fitting DGPs for Marijuana legalization. The figure depicts DGPs with KS distance < .005 (from 1,000,000 simulations) for the empirical case study. Each point represents an accepted draw. We rounded the balance parameters to show conditional relationships.

These patterns make sense when we examine the conditional relationships among our parameters, presented in Figure 4. Each panel shows a particular balance value assumed in the simulations. The *x*-axis shows the parameter values for rate of change, while the y-axis shows the parameter values for strength. Each point thus represents a particular DGP with a known rate of change, strength of change, and balance, which "survives" our ABC procedure, thereby representing a good fit for the data. We see that the algorithm can easily suggest, at least for this particular case, a specific direction estimate—median balance of .82 with an IQR of .75 and .90—but it converges to almost analytically determined bands in high-balance panels: we detect either small changes in a large number of people, or large changes in a small number of people. To put these results in perspective, note that the KS distance among these accepted draws is virtually indistinguishable from one another, most having a difference less than .001, thus having equal consistency with the observed data.

There is one strong implication of this exercise: there are competing generative models that may easily explain the same aggregate patterns, and there is no methodological trick to solve this issue. If we take the measurement error problem to its natural conclusions, the same data can be explained by incompatible theoretical mechanisms, and our conclusions may be wildly off-the-mark.

Exploring the case of the legality of abortion

What about cases where the direction or even existence of change is less clear-cut? We now turn our attention to the legality of abortion debate, a contentious issue characterized by high

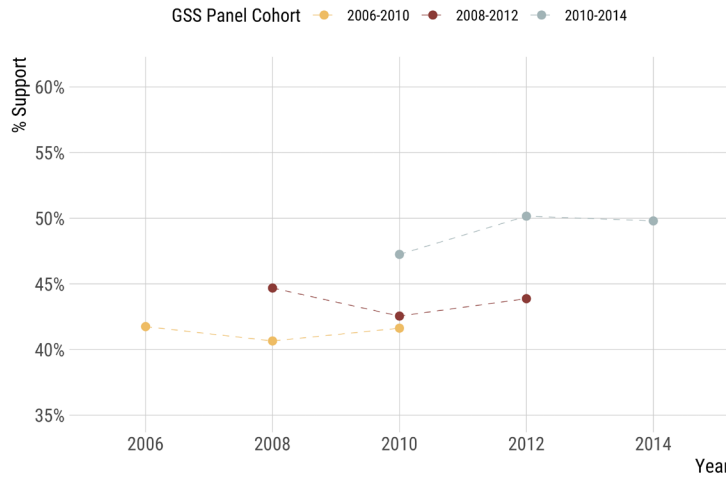

FIGURE 5 The percent support for the legality of abortion. The figure depicts the percent of American adults saying "it should be possible for a pregnant woman to obtain a legal abortion if the woman wants it for any reason" in the General Social Survey 2006, 2008, and 2010 panel studies. We included respondents who participated in all three waves of each panel study and weighted the results using GSS post-stratification weights. Note that panel attrition caused compositional changes in the estimates, and compared to the GSS time-series data, the estimates of support are marginally inflated.

disagreement and high political sorting in the American context (Baldassarri & Gelman, 2008; Hout et al., 2022). In 2006, 2008, and 2010 GSS panels, respondents agreed or disagreed with the statement: "it should be possible for a pregnant woman to obtain a legal abortion if the woman wants it for any reason." Figure 5 shows response trajectories for this question across three panels, indicating a clear stasis.

We apply our grid-search algorithm once more, using the same specifications we applied in the case of marijuana legalization: 1,000,000 draws with $R \in [0, 1]$, $S \in [0, 2]$, and $D \in [0, 1]$, with reliability score ranging between 60% and 100%, each having uniform priors. This time, we kept DGPs with KS distance < .0025 given the high number of "better" fits we observed, resulting in 1534 DGPs. Figure 6 shows the pair plots, while Figure 7 shows the conditional estimates. Once again, we repeat the pair plot logic in our presentation: diagonal panels in Figure 6 display the marginal posterior for each parameter, while off-diagonal panels show how accepted draws cluster. Similarly, we fix balance values in Figure 7 and trace the plausibility bounds for rate and strength.

We find strong variability in the plausible generative scenarios. Note how, in all cases, the accepted draws are squished toward 0—either via 0 rate of change or via 0 strength of change—indicating a clear preference toward no change in the underlying processes. That said, the plausibility bounds also become strongly noisy when the estimated balance is even: accepting scenarios ranging from no change to the now-familiar high change in a low number of people and low change in a high number of people. The IQR of KS distance among all accepted samples is between .001 and .002, suggesting that model fits are reasonably similar across all scenarios.

Exploring the case of immigration spending

These exercises are pertinent illustrations of the limitations of panel studies. Based on the observed data, we cannot determine whether the observed marginals result from a small number of people making substantial belief changes or a large number of people making

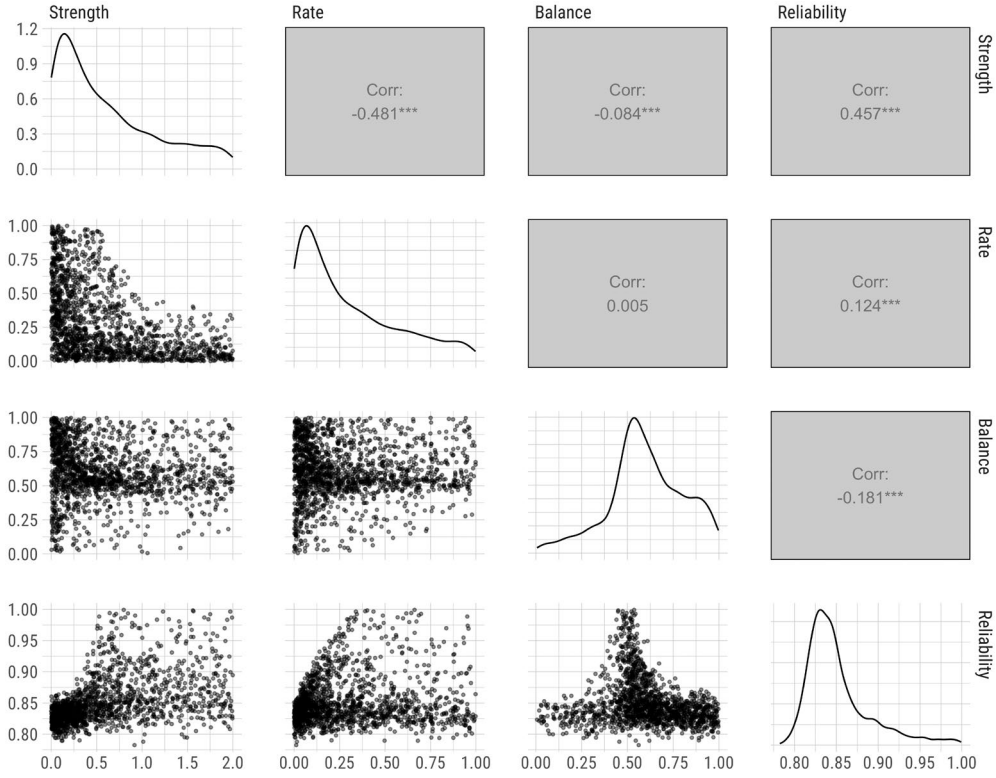

FIGURE 6 Pair plots for the legality of abortion. The figure depicts the distribution of, and bivariate relationships among, grid-search parameter values from the best-fitting DGPs for the case of the legality of abortion (KS distance < .0025).

minor political changes. This may, ultimately, boil down to the issue of observation frequency: with three-wave panel data and reasonable doubts about the accuracy of every single survey response due to reliability and resolution concerns, it is challenging to adjudicate the underlying generative models. We now relax this assumption and extend our method to cases with more than three waves.

The data for this exercise comes from the *Political Psychology Data from a 26-Wave Yearlong Longitudinal Study* (Brandt et al., 2021), conducted biweekly in 2019–2020 for 1 year. In one question, respondents were asked whether they agree with the policy of increased federal spending on immigration control. We binarized the responses to this question such that those agreeing with the statement received a 1, and those who said the spending should be decreased or kept the same received a 0. We then dropped respondents with missing data and trimmed the periods to pre-COVID windows, resulting in 213 individuals observed 17 times over 34 weeks. Although the sample for this case study is obviously not representative, it provides an intensive set of repeated observations at the individual level (see Figure 8).

We applied the same procedure, but we found that our KS distance cutoff of .005 was too low—we ended up with no accepted samples, suggesting that the model simply cannot converge to a reasonably strong fit for this case study (we discuss the implications of this finding below). We then redefined the cutoff point as a KS distance of .05, ending up with 3676 accepted samples.

Once again, Figure 9 presents the pair plots, while Figure 10 depicts the conditional DGP estimates. Consistent with the decline in the observed marginals, we find accepted samples in the regions with balance lower than .5 (a median of .14 with an interquartile range between .07

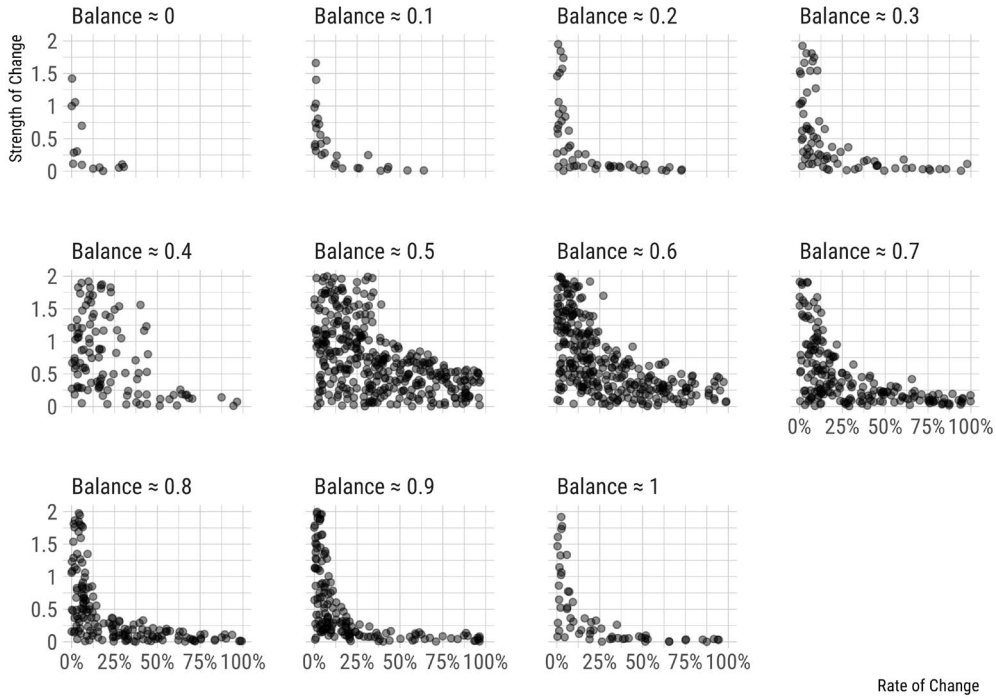

FIGURE 7 Best-fitting DGPs for the legality of abortion. The figure depicts DGPs with KS distance < .0025 (from 1,000,000 simulations) for the empirical case study. Each point represents an accepted draw. We rounded the balance parameters to show conditional relationships.

and .22). We also find similar IQR values of .35 for strength and .37 for rate, suggesting plausibility bounds in our conditional estimates. In contrast to the marijuana legalization case, however, these plausibility bounds have much more noise, as seen in Figure 10, suggesting that the data has many plausible DGPs spread across the grid, without a clear band forming in our accepted samples.

The detection of true changers

So far, we have presented evidence for identifying plausible DGPs that may have produced the observed empirical data. But what about classifying individuals into those who change and those who do not, that is, knowing *who changed*? Can we also recover an accurate classification of changers and non-changers at the individual level of observation (Tharwat, 2020)?

To address this question, we conduct a simulation exercise. For each of our empirical case studies, we first generate 1000 datasets by fixing the parameters of interest to observed and assumed values. For the case of marijuana legalization, for instance, we fix the number of individuals, the number of waves, and the balance of the marginals to the observed GSS values, while we fix the balance and reliability scores to the median of our grid-search results and the rate of change and strength of change to a plausible scenario for 50% and 1 SD, respectively. Similarly, for the legality of abortion and immigration spending, we assumed a 20% rate with 2 SD strength as well as a 100% rate with .25 SD strength, while fixing everything else to observed values in the data or median values. These scenarios are reasonable fits, but for each simulated dataset, we *know* who the real changers are.

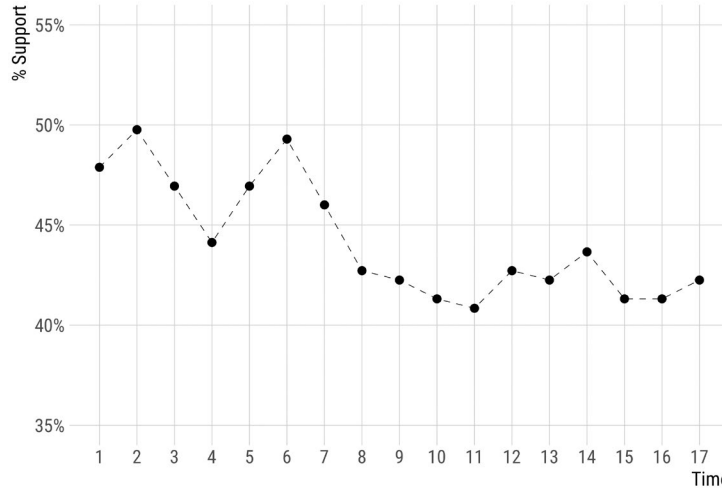

FIGURE 8 The percent support for an increase in the control of immigration spending. The figure depicts the percent support for the statement "the federal spending to control immigration should be increased" in the Political Psychology Data from a 26-Wave Yearlong Longitudinal Study.

Our metric for classification success is quite straightforward. We first use the "observed" simulation data to fit individual-level regression models, similar to our grid-search procedure:

$$log\left(\frac{\mathbb{P}(y=1)}{1-\mathbb{P}(y=1)}\right) = \beta_0 + \beta_1 t, t \in [0,1]$$

where t is a normalized time variable. We then estimate a predicted change score by calculating the difference between the prediction at $t = 1$ and the prediction at $t = 0$.

Once we have all the change scores, we rank individual cases from highest to lowest, and based on our true DGP parameters, select respondents as "changers." For instance, if we assumed 70% positive change and 30% negative change, we select the highest 70% of individual slopes as positive changers and the lowest 30% of individual slopes as negative changers. In the last step, we examine whether we can recover true changers in our DGPs from these scores.

Figure 11 shows the distribution of accuracy scores across all studies, each having 1000 simulations for change detection. The median accuracy score is 43% for the case of marijuana legalization, 36% for the legality of abortion, and 61% for immigration spending, meaning that among all true changers, we can only correctly classify 36%–61% of them, even with full knowledge of the DGPs.

THEORETICAL AND EMPIRICAL IMPLICATIONS

We now examine the main conclusions of our case studies and draw several implications for the use of panel data in studies of political change and stability across the life course.

1. Our case studies highlight that the aggregate change patterns we observe may emerge from vastly different compositional processes. This implies that researchers should be cautious when interpreting the results from panel studies, and think carefully about the measurement problem. To bolster this point, we provide two additional studies in Supporting Information A—the first case looking at a tumultuous social period during the COVID-19 pandemic and examining patterns of "generalized trust" (using data

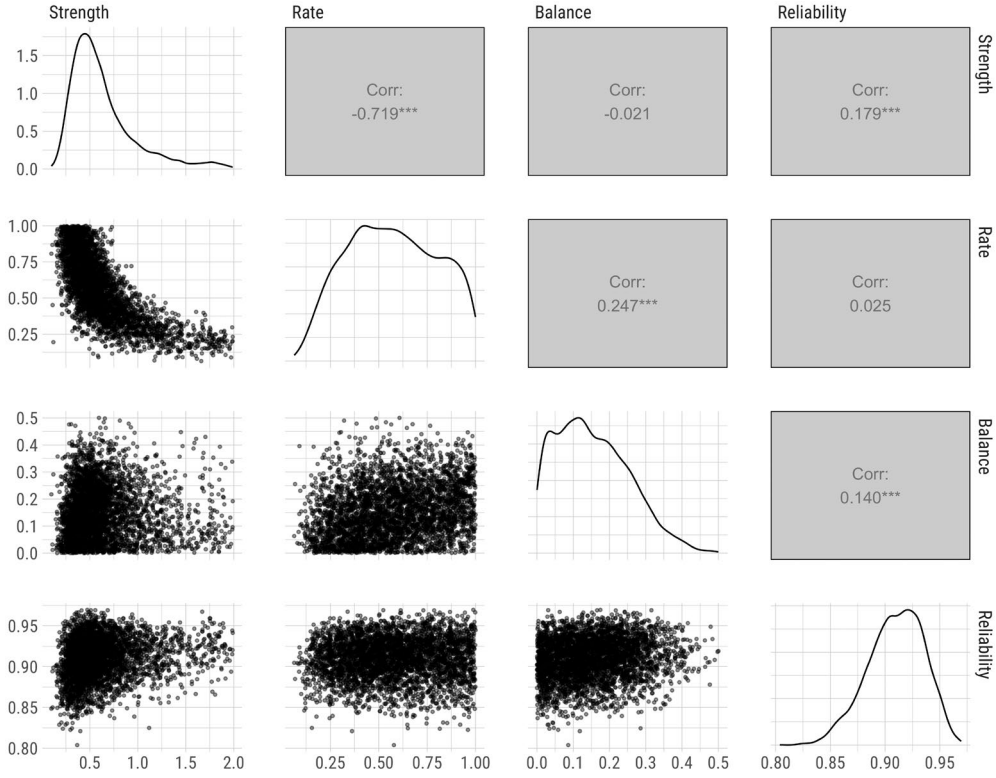

FIGURE 9 Pair plots for immigration spending. The figure depicts the distribution of, and bivariate relationships among, grid-search parameter values from the best-fitting DGPs for the case of immigration spending (KS distance < .05).

from the British Election Study panel), and the second case looking at adolescents over a 10-year window to explore "belief in God" (using data from the National Study of Youth and Religion).

2. We emphasize that this identification problem cannot be "solved" with simple methodological considerations. As we emphasized before, our objective was not to eliminate the problem, but rather to provide a concrete approach that might help us understand the issue more clearly. More substantively, we need to acknowledge the problem in our empirical applications and be willing to *impose* certain assumptions. The fact that multiple—sometimes contradictory—generative processes may very well account for the same data means that researchers should assert theoretical positions when it comes to connecting their hypotheses to empirical data.
3. Our studies present interesting variations across (a) model fit, that is, whether our grid-search approach can get reasonably close to true DGPs, and (b) the noise around plausibility bounds. To understand these dynamics better, we conducted a simulation exercise and explored the effects of reliability and the number of waves on model fit (see Supporting Information B). We show that an increased reliability score and a higher number of waves mean increased model fit—the effects of which depend on the underlying real-world parameters. This suggests that a better fit for plausible DGPs requires researchers to collect as many waves as possible and a strong investment in reliable response measurement strategies. Without these, researchers will need strong theoretical assumptions to narrow down plausible data generating mechanisms.

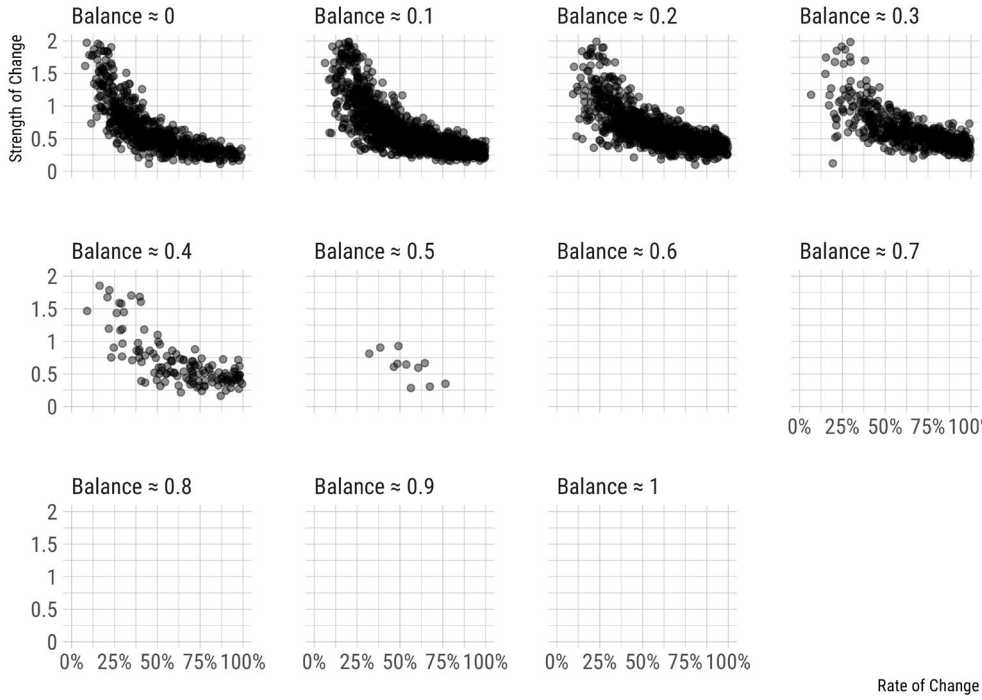

FIGURE 10 Best-fitting DGPs for immigration spending. The figure depicts DGPs with KS distance < .05 (from 1,000,000 simulations) for the empirical case study. Each point represents an accepted draw. We rounded the balance parameters to show conditional relationships.

FIGURE 11 The distribution of accuracy scores across empirical case studies. The figure depicts the distribution of accuracy scores for the detection of changers across three empirical case studies. For each case, we performed 1000 simulations for the classification task.

4. We showed that accurately identifying changers at the individual level might be challenging. In Supporting Information C, we present findings from a simulation study, which explores 360 DGPs across a wide variety of scenarios—varying rate, strength, directional balance, reliability, and the number of waves. We show that accuracy scores range from 8% to 74%, with more reliability and a higher number of waves significantly increasing our ability to classify individuals correctly. Once again, these findings suggest that researchers should

strive to increase the reliability of response measurement and try to collect as many study waves as possible in order to have a decent shot at identifying change at the individual level.[6]

5. For applied researchers, we recommend our diagnostic workflow. The detection of an equivalence class of generative models means that we can have a transparent and concrete checklist for substantive inference. To do so, researchers should identify plausible generative processes that may have generated their data, and either note this heterogeneity or explicitly *assume* a certain process is at play before moving forward with their substantive interpretations.

DISCUSSION AND CONCLUSIONS

In this article, we examined the identification of belief change in a panel context. Using simulations and empirical case studies, we showed that a grid-search procedure that iterates over generative models can help us identify plausible data generation processes that may have produced our empirical observations, though the identification of individual cases occurs under narrow conditions. We argued that the *group-to-person generalizability problem* is an impediment to a robust understanding of political belief change, and scholars should understand the extent to which their target estimates may generalize to the population of their study. As such, we proposed a simulation approach that takes response generation seriously, which may help us move forward in addressing this problem.

This article has several implications for the study of political belief change across the life course. First, our findings suggest that recent debates about the existence or non-existence of belief change in adult populations (Kiley & Vaisey, 2020; Lersch, 2023) should tackle the heterogeneous nature of change and stability in individual preferences because group-level measures might hamper our understanding of individual mechanisms. Instead, scholars need to specify the individual-level generative models that aggregate average estimates imply. Just like analyses that examine population patterns to adjudicate compositional and individual changes (Bartels & Jackman, 2014; Vaisey & Lizardo, 2016), we need robust and reliable panel studies to identify the underlying generative processes.

Second, the ambivalence in response trajectories indicates that classical solutions to the measurement problem—ranging from scale construction to structural equation models and principal component analyses (Alwin & Krosnick, 1991; Ansolabehere et al., 2008; Judd & Milburn, 1980)—should have an important place in panel designs, as even moderate reliability scores can significantly restrict our ability to specify generative processes from observed data.

One theoretical assumption of this article was to rely on a latent variable model, and avoid questions about contradictory considerations in human cognition that mitigate stable and coherent opinion formation (Converse, 1964; Zaller & Feldman, 1992). This issue, of course, introduces another layer to the problem of belief change and measurement error. We relied on a very strong assumption that an individual has a central tendency when it comes to political issues, rather than being functionally "random" in their responses to survey questions. However, a careful examination of adult belief change should be grounded in a more theoretically sound study of belief formation.

[6]In Supporting Information D, we show additional analyses comparing our fixed-effects classification approach with multilevel mixed models (Lersch, 2023). We found that a fixed-effects approach largely fares better, even though a mixed model incorporates information from the overall sample distributions. This, however, largely reflects the fact that we have not incorporated a mixture process: a mixture mixed effects model or a latent transition model might have allowed us to improve these estimates. Since, for the purposes of this paper, we focused on the group-to-person generalizability problem, we bracket this issue from further consideration, but we hope future research will improve this strategy.

Finally, these results call for granular panel surveys that go beyond the classical 2-wave and 3-wave designs because reliability concerns significantly limit the plausible conclusions we may derive from these shorter panels. In a 2-wave setup, where there is only one change, we can never properly establish whether a change we observe is due to measurement error or it is indeed a genuine shift, making extended panel designs necessary. Since we care about generative models rather than one observed sample, strategies that enable us to reliably estimate target quantities are warranted and essential to make progress on longstanding questions about belief change.

ACKNOWLEDGMENTS

We thank the participants of the *Worldview Lab at Kenan Institute for Ethics* at Duke University, as well as sociology colloquium participants at New York University and the University of Chicago for their insightful and helpful comments.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in Open Science Framework at https://osf.io/5x48d/.

ORCID

Turgut Keskintürk https://orcid.org/0000-0002-0270-0006

REFERENCES

Achen, C. H. (1975). Mass political attitudes and the survey response. *American Political Science Review, 69*(4), 1218–1231. https://doi.org/10.2307/1955282

Allison, P. D. (2009). *Fixed effects regression models*. SAGE Publications.

Alwin, D. F. (2007). *Margins of error: A study of reliability in survey measurement*. John Wiley & Sons.

Alwin, D. F., & Krosnick, J. A. (1991). The reliability of survey attitude measurement: The influence of question and respondent attributes. *Sociological Methods & Research, 20*(1), 139–181. https://doi.org/10.1177/0049124191020001005

Ansolabehere, S., Rodden, J., & Snyder, J. M. (2008). The strength of issues: Using multiple measures to gauge preference stability, ideological constraint, and issue voting. *American Political Science Review, 102*(2), 215–232. https://doi.org/10.1017/S0003055408080210

Baldassarri, D., & Gelman, A. (2008). Partisans without constraint: Political polarization and trends in American public opinion. *American Journal of Sociology, 114*(2), 408–446. https://doi.org/10.1086/590649

Bartels, L. M., & Jackman, S. (2014). A generational model of political learning. *Electoral Studies, 33,* 7–18. https://doi.org/10.1016/j.electstud.2013.06.004

Beaumont, M. A. (2010). Approximate Bayesian computation in evolution and ecology. *Annual Review of Ecology, Evolution, and Systematics, 41*(1), 379–406.

Brandt, M. J., Turner-Zwinkels, F. M., & Kubin, E. (2021). Political psychology data from a 26-wave yearlong longitudinal study (20192020). *Journal of Open Psychology Data, 9*(1), 2. https://doi.org/10.5334/jopd.54

Converse, P. E. (1964). The nature of belief systems in mass publics. In D. Apter (Ed.), *Ideology and discontent* (pp. 206–261). Princeton University Press.

Fosse, E., & Winship, C. (2019). Analyzing age-period-cohort data: A review and critique. *Annual Review of Sociology, 45*(1), 467–492.

Hout, M., & Hastings, O. P. (2016). Reliability of the Core items in the general social survey: Estimates from the three-wave panels, 20062014. *Sociological Science, 3,* 9711002. https://doi.org/10.15195/v3.a43

Hout, M., Perrett, S., & Cowan, S. K. (2022). Stasis and sorting of Americans' abortion opinions: Political polarization added to religious and other differences. *Socius: Sociological Research for a Dynamic World, 8,* 237802312211176. https://doi.org/10.1177/23780231221117648

Judd, C. M., & Milburn, M. A. (1980). The structure of attitude systems in the general public: Comparisons of a structural equation model. *American Sociological Review, 45*(4), 627–643. https://doi.org/10.2307/2095012

Kiley, K., & Vaisey, S. (2020). Measuring stability and change in personal culture using panel data. *American Sociological Review, 85*(3), 477–506. https://doi.org/10.1177/0003122420921538

Lersch, P. M. (2023). Change in personal culture over the life course. *American Sociological Review, 88*(2), 252–283. https://doi.org/10.1177/00031224231156456

McManus, R. M., Young, L., & Sweetman, J. (2023). Psychology is a property of persons, not averages or distributions: Confronting the group-to-person generalizability problem in experimental psychology. *Advances in Methods and Practices in Psychological Science, 6*(3), 1–23. https://doi.org/10.1177/25152459231186615

Ochoa, N. R., & Vaisey, S. (2024). Opinions on hard-to-discuss topics change more via cohort replacement. *Evolutionary Human Sciences, 6,* e25. https://doi.org/10.1017/ehs.2024.13

Owens, L. A., & Pedulla, D. S. (2014). Material welfare and changing political preferences: The case of support for redistributive social policies. *Social Forces, 92*(3), 1087–1113.

Rüttenauer, T., & Ludwig, V. (2020). Fixed effects individual slopes: Accounting and testing for heterogeneous effects in panel data or other multilevel models. *Sociological Methods & Research, 52*(1), 43–84. https://doi.org/10.1177/0049124120926211

Ryder, N. B. (1965). The cohort as a concept in the study of social change. *American Sociological Review, 30*(6), 843–861. https://doi.org/10.2307/2090964

Sisson, S. A., Fan, Y., & Beaumont, M. (2018). *Handbook of approximate Bayesian computation*. CRC Press.

Small, M. L. (2002). Culture, cohorts, and social organization theory: Understanding local participation in a Latino housing project. *American Journal of Sociology, 108*(1), 1–54.

Swidler, A. (1986). Culture in action: Symbols and strategies. *American Sociological Review, 51*(2), 273–286. https://doi.org/10.2307/2095521

Tharwat, A. (2020). Classification assessment methods. *Applied Computing and Informatics, 17*(1), 168–192. https://doi.org/10.1016/j.aci.2018.08.003

Vaisey, S., & Kiley, K. (2021). A model-based method for detecting persistent cultural change using panel data. *Sociological Science, 8,* 83–95. https://doi.org/10.15195/v8.a5

Vaisey, S., & Lizardo, O. (2016). Cultural fragmentation or acquired dispositions? A new approach to accounting for patterns of cultural change. *Socius: Sociological Research for a Dynamic World, 2,* 1–15. https://doi.org/10.1177/2378023116669726

Vaisey, S., & Miles, A. (2017). What you Canand Can'tdo with three-wave panel data. *Sociological Methods & Research, 46*(1), 44–67. https://doi.org/10.1177/0049124114547769

Zaller, J., & Feldman, S. (1992). A simple theory of the survey response: Answering questions versus revealing preferences. *American Journal of Political Science, 36*(3), 579. https://doi.org/10.2307/2111583

Zaller, J. R. (1992). *The nature and origins of mass opinion*. Cambridge University Press.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Keskintürk, T., Bello, P., & Vaisey, S. (2026). The promises and pitfalls of using panel data to understand individual belief change. *Political Psychology*, *47*, e70056. https://doi.org/10.1111/pops.70056