Political Analysis Advance Access published May 7, 2015

Political Analysis (2015) pp. 1–27
doi:10.1093/pan/mpv011

Explaining Systematic Bias and Nontransparency in U.S. Social Security Administration Forecasts

Konstantin Kashin and Gary King
Institute for Quantitative Social Science, Harvard University, Cambridge, MA 02138

Samir Soneji
The Dartmouth Institute for Health Policy & Clinical Practice, Dartmouth College, Lebanon, NH 03766
e-mail: king@harvard.edu (corresponding author)

Edited by R. Michael Alvarez

The accuracy of U.S. Social Security Administration (SSA) demographic and financial forecasts is crucial for the solvency of its Trust Funds, other government programs, industry decision-making, and the evidence base of many scholarly articles. Because SSA makes public insufficient replication information and uses antiquated statistical forecasting methods, no external group has ever been able to produce fully independent forecasts or evaluations of policy proposals to change the system. Yet, no systematic evaluation of SSA forecasts has ever been published by SSA or anyone else—until a companion paper to this one. We show that SSA's forecasting errors were approximately unbiased until about 2000, but then began to grow quickly, with increasingly overconfident uncertainty intervals. Moreover, the errors are largely in the same direction, making the Trust Funds look healthier than they are. We extend and then explain these findings with evidence from a large number of interviews with participants at every level of the forecasting and policy processes. We show that SSA's forecasting procedures meet all the conditions the modern social-psychology and statistical literatures demonstrate make bias likely. When those conditions mixed with potent new political forces trying to change Social Security, SSA's actuaries hunkered down, trying hard to insulate their forecasts from strong political pressures. Unfortunately, this led the actuaries into not incorporating the fact that retirees began living longer lives and drawing benefits longer than predicted. We show that fewer than 10% of their scorings of major policy proposals were statistically different from random noise as estimated from their policy forecasting error. We also show that the solution to this problem involves SSA or Congress implementing in government two of the central projects of political science over the last quarter century: (1) transparency in data and methods and (2) replacing with formal statistical models large numbers of ad hoc qualitative decisions too complex for unaided humans to make optimally.

1 Introduction

Social Security is the single largest program in the U.S. government, currently providing benefits to over 58 million retirees and disabled and levying payroll taxes on another 210 million workers. It is also one of the most popular programs, ending retirement-generated impoverishment for a vast segment of the population. The program functions via an intergenerational transfer of wealth. The government levies payroll taxes on today's workers and deposits the revenue into interest-earning Social Security Trust Funds; retirees, disabled workers, and their families receive benefits paid from the Trust Funds. The Trust Funds function as a bank account that enables the Social Security Administration (SSA) to smooth out temporary imbalances between workers and beneficiaries. For example, while the baby boom generation remains in the workforce, the Trust Funds collect more

**Authors' note:* For helpful advice or comments, we are grateful to Bill Alpert, Jim Alt, Steve Ansolabehere, Neal Beck, Nicholas Christakis, Mo Fiorina, Dan Gilbert, Alexander Hertel-Fernandez, Martin Holmer, David Langer, and Theda Skocpol. Thanks also to the many participants in the forecasting and policy process for information and advice. Replication data are available on the *Political Analysis* Dataverse at http://dx.doi.org/10.7910/DVN/28323. Supplementary materials for this article are available on the Political Analysis Web site.

money than SSA needs to pay for current beneficiaries. When this generation retires, it withdraws more benefits from the Trust Funds than future workers' payroll taxes will generate.

The viability of Social Security depends fundamentally on the solvency of the Trust Funds, which, in turn, depends upon accurate demographic and financial forecasts that enable members of Congress to make sound decisions on Social Security policy. For example, without adjustments to payroll taxes or benefit levels, the Trust Funds could become exhausted sooner than anticipated if medical advancements extend the life expectancy of retirees or an economic recession increases unemployment and decreases payroll tax revenue.

The goal of SSA's demographic and financial forecasts is to provide sufficiently accurate information to policymakers to address excess cash outflows or low cash inflows via one or more of the available policy levers. Among many others, these include gradual increases in payroll tax rates or changes in the retirement age. The earlier an accurate forecast becomes available, the more options Congress has to ensure the solvency of Social Security through incremental, politically palatable, and economically achievable changes. In contrast, inaccurate demographic and financial forecasts narrow the range of feasible policy levers, often to proposals that are fiscally disruptive, politically challenging, or otherwise infeasible.

The errors in SSA forecasts are large and mostly in the same direction, making the Trust Fund look healthier than it actually is. With these forecast evaluations, we also provide the first honest uncertainty estimates of SSA's policy scores, which constitute the only real evaluation of every major proposal to change the system in the past two decades; we find that more than 90% of SSA's numbers are overwhelmed by forecast uncertainty.

Of course, even the most skilled forecasters will sometimes make inaccurate predictions due to unforeseeable events. Such surprises do not reflect negatively on the forecasters. However, continuing to make systematic forecasting errors is evidence of bias or flawed methodology. In this case, it suggests a failure to meet the "best practices" in scientific evaluation procedures commonplace in academia, industry, and other government agencies. In fact, we find that the SSA lags behind best practices benchmarks in at least three critical respects.

First, SSA has never published systematic and comprehensive evaluations of its forecasts. Second, the SSA Office of the Chief Actuary (OCACT), which produces the forecasts, withholds many aspects of its data and forecasting procedures from the public, the scientific community, and even other parts of SSA. This makes independent replication impossible. Third, critical aspects of OCACT's forecasting procedures are informal or qualitative, even though far better systematic quantitative techniques exist and continually improve. Consequently, SSA is not set up to learn optimally from its pattern of forecasting errors, or to meet the "replication standard" widely supported in academia (King 1995) and even in the U.S. government via President Obama's executive orders (j.mp/ObamaOpenData). As a result, important parts of OCACT procedures are difficult or impossible to replicate, understand, or implement, which means that errors introduced are unlikely to be corrected. Indeed, some steps in SSA's forecasting involve committees or individuals making large numbers of interrelated qualitative judgments that are extremely difficult to do well without computer assistance and, at the same time, involve high levels of discretion for decision makers hidden from public view.

As we show here, SSA's forecasting procedures turn out to meet all the major conditions for unintended political, social, and psychological biases to be introduced, even when those involved try diligently to produce forecasts free from external influence. Our research seems to indicate, consistent with the social-psychological literature, that these biases do not occur because of individuals making avoidable mistakes or not trying hard enough. Greater effort for the same task performed in the same manner would likely not help. Rather, the biases occur because of the lack of formal procedures at SSA designed to avoid these biases in the first place.

Indeed, OCACT's own scientific panel of distinguished outside advisers, the Social Security Advisory Board's Technical Panel on Assumptions and Methods, has frequently recommended that OCACT make data and replication procedures available and improve their statistical methodology and uncertainty estimation. We demonstrate here that SSA regularly ignores this and many other recommendations of its scientific advisory panel, and this oversight provides important evidence on the evolution of biases inherent in SSA forecasting.

Section 2 outlines some of the methodological challenges involved in Social Security forecasts and then summarizes and extends our evidence from Kashin, King, and Soneji (2015c) on the biases in SSA's forecasts. Section 3 offers a hypothesis about how these systematic biases came about, supported by research from social psychology on situations like these where unnecessary and uncorrected biases are generated.

To gather information and evaluate our theories, we conducted a large number of semi-structured personal interviews with those in, and involved with, SSA at every level of government and the policy process. Since the politics of Social Security has become extremely polarized and highly conflictual, with SSA administrators deeply involved in many aspects of it, we keep the identities of our interviewees confidential. With this information, Section 4 summarizes the pressures on SSA that likely led to these forecasting biases. In order to reduce the possibility of bias going forward, Section 5 provides an overview of methods that formalize ad hoc, qualitative forecasts. Section 6 concludes. An appendix gives details of the data used in this article, with a complete replication data set available at Kashin, King, and Soneji (2015a).

2 Social Security Administration Forecasting

In this section, we briefly summarize and extend some of the empirical results from Kashin, King, and Soneji (2015c) about the performance of SSA's forecasting, convey the methodological challenges in the complex data used for Social Security forecasting, and outline some of the specific procedures SSA uses to forecast.

2.1 *Social Security Forecasting Performance*

We summarize our prior work in two points. First, despite many years of approximately unbiased forecasting, OCACT began issuing systematically biased forecasts, with overconfident assessments of uncertainty, after about the year 2000. Instead of following best practices and learning from these biases to improve subsequent forecasts, the biases have grown much larger over time. These biases led members of Congress, other policymakers, and the public to conclude each year that the Social Security Trust Funds would be on firmer financial footing than actually turned out to be the case, year after year.

Second, OCACT does not share all its data, code, and replication information with the public, the scientific community, and other parts of SSA or the federal government. We also discuss here the contributing factor that OCACT's forecasting procedures are informal and qualitative, and fail to take advantage of the dramatic improvements in statistical modeling, and ways of formalizing informal procedures like these, that have been developed over the past quarter century.

2.2 *Methodological Challenges*

The statistical challenges in accurately forecasting Social Security are substantial. They involve modeling complex multivariate data structures and incorporating numerous details in modeling how this enormous and complicated government program is administered, run, and funded. We focus here on the crucial methodological challenge of forecasting mortality rates—a key input to SSA's financial forecasts. An accurate forecast of mortality is necessary to estimating the number of workers contributing payroll taxes and the number of retirees and disabled receiving benefits. To begin to provide a feel for the problem, consider Fig. 1.

The key to this figure is simultaneously recognizing both the methodological challenges and the substantial information available to improve forecasting accuracy using the right model. First, consider the top row of Fig. 1, which shows the probability of a male (top left panel) or female (top right panel) dying within 1 year, for each year of age (0–109, colored from red for the youngest ages to purple for the oldest ages) and for every calendar year from 1980 to 2010 along the horizontal axis. Small changes in any one line may contribute to large changes in the number of workers or the number of retirees. Thus, the highly regular aggregate patterns also contain important and apparently small variations that need to be modeled carefully. Each line in each graph is a

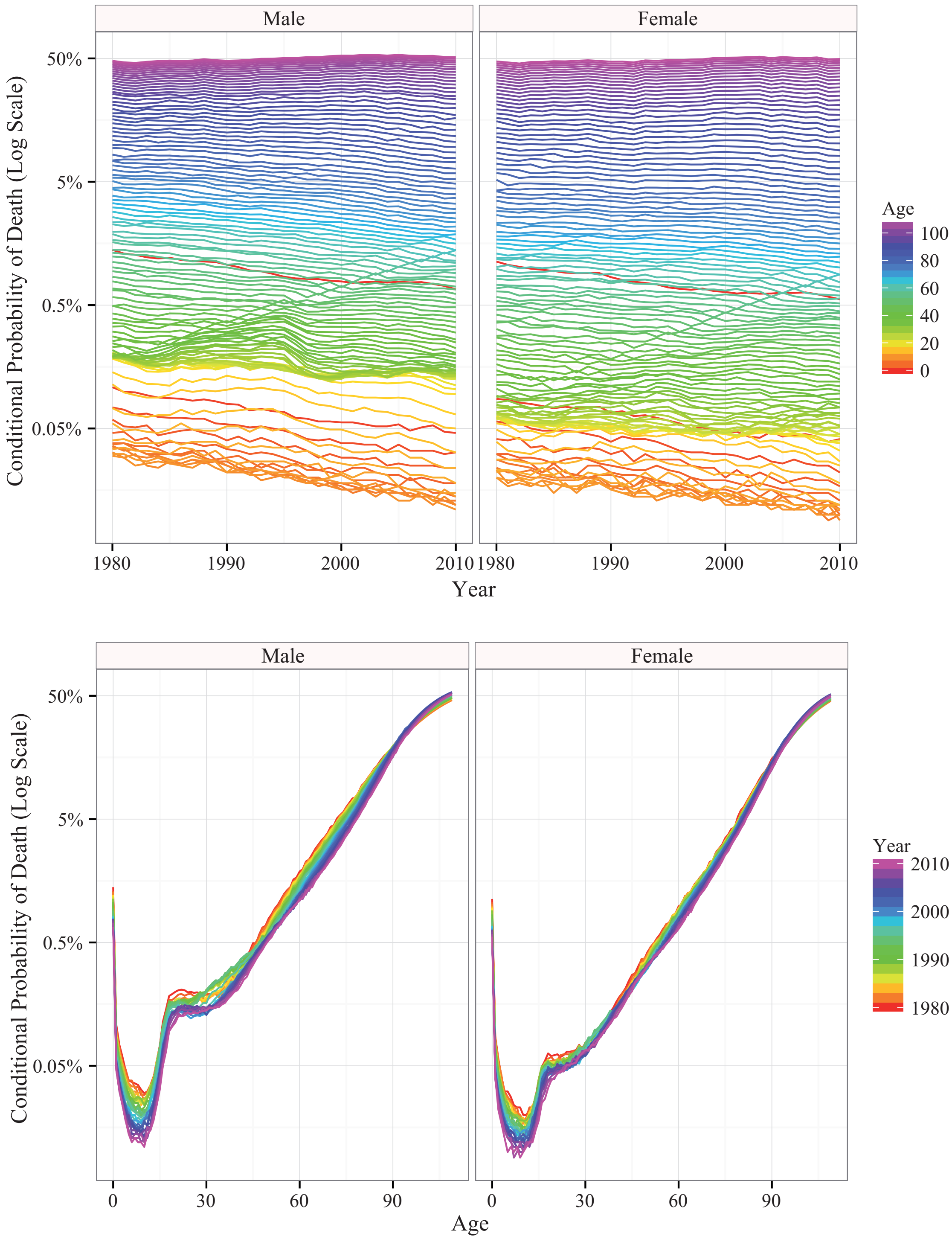

Fig. 1 Time and age profiles of conditional probabilities of death (Human Mortality Database).

time-series plot representing different cohorts of people at the same ages at different times. One can also see important diagonal patterns (from bottom left to top right in each of the top panels) representing the continued experience of higher or lower mortality of the same birth cohort as it ages over time.

These two top panels in Fig. 1 convey two other crucial facts that need to be taken into account in any serious model of the data-generation process. First, although mortality is relatively smooth

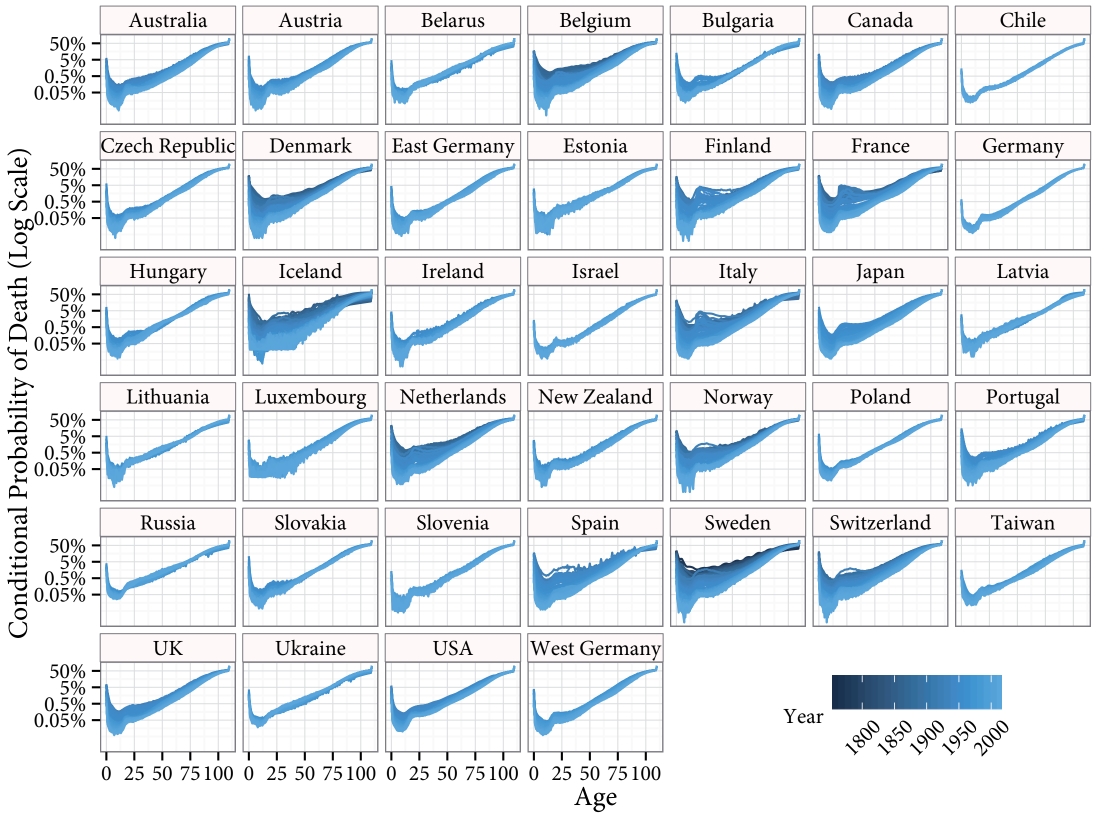

Fig. 2 Conditional probabilities of death (Human Mortality Database).

over time, forecasting any one of these lines would be less certain many years into the future. Second, there is considerable information in adjacent age groups that can be used to improve the forecasts if used appropriately. This second point can be seen even more clearly in the bottom two panels of the same figure, which re-express the same data with age along the horizontal axis and colors for each year. So, instead of time-series plots, these panels portray age profiles. The characteristic shape is common in demographic data across countries and time periods, indicating that the few years after birth are relatively risky, after which mortality drops, and then starting at about age 5–10 years mortality inexorably increases, almost log-linearly, except for a bump that coincides with higher risk of accidental mortality among older adolescents and young adults.

The opportunities in modeling mortality involve recognizing the powerful information available to build into forecasting methods. In fact, the patterns in Fig. 1 turn out not to be unique to this time period or even the United States and so should be considered valuable information for model building, such as for constructing Bayesian priors. To convey how stable these patterns are, we constructed Fig. 2, which gives the log-mortality age profile for 39 separate countries (lighter blue indicates later years). As is apparent, the general pattern from U.S. age profiles holds with remarkable generality across countries. Clearly, forecasting methods that ignore or do not formally encode this powerful information are, at best, highly inefficient.

The patterns in Figs. 1 and 2 represent highly reliable demographic knowledge that forecasting methods should include. This information has been formally encoded in priors by smoothing across expected mortality in neighboring age groups and adjacent time points and their interactions (Girosi and King 2008; King and Soneji 2011), or via principal component modeling (Lee and Carter 1992). However, instead of sophisticated statistical methods encoding this well-known demographic knowledge, SSA manually adjusts simple regression models of mortality on time to be consistent with the views of a committee of actuaries making qualitative judgments. Important covariates such as smoking and obesity are formally excluded from their statistical model, although they are considered qualitatively.

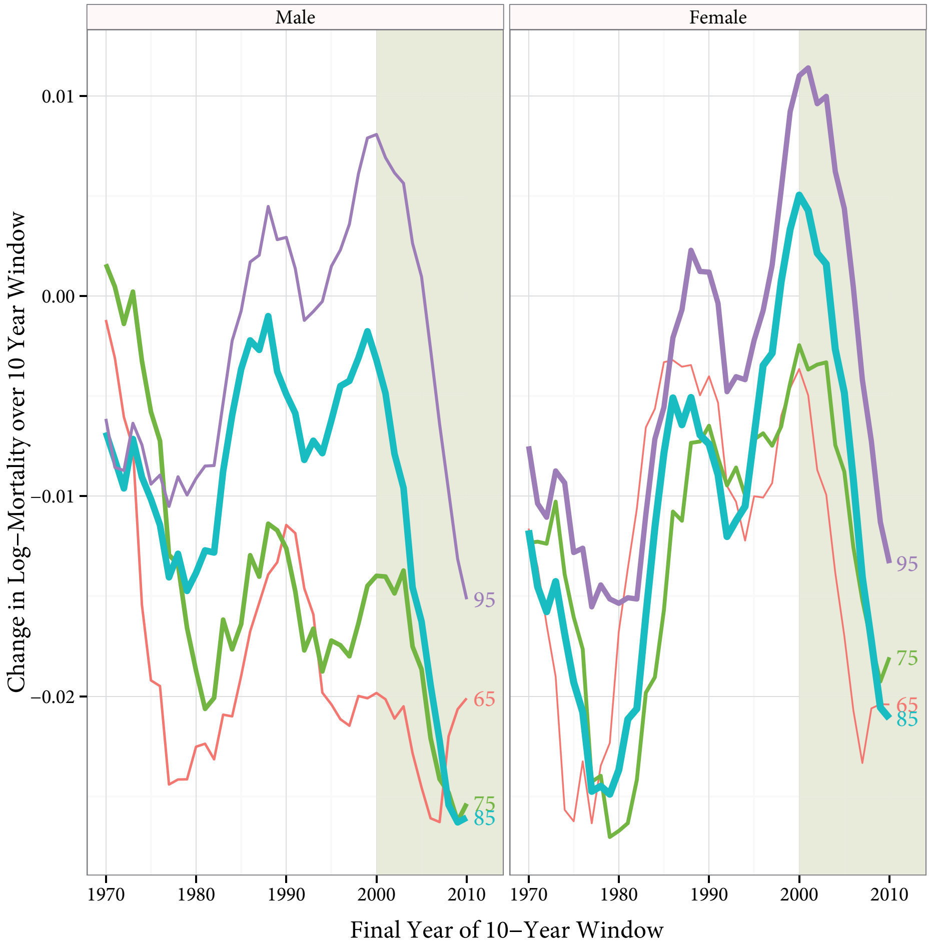

Fig. 3 Changes in log-mortality over successive 10-year windows. Each line is an age group, with the thickness of the line drawn proportional to the number of deaths in that age group. The confidence interval around each of the lines (which we do not add to the graph for clarity) is approximately ±0.004.

Finally, we summarize patterns in log-mortality changes to illustrate a key feature of U.S. mortality data that will prove important in explaining the patterns of bias in SSA forecasts. We focus on age groups 65 years and older, most of whom will be drawing Social Security benefits for as long as they are alive. To do this, we compute the change in log-mortality over successive 10-year time intervals for selected age groups. We thus regress log-mortality on time, with observations from t–9 to t (repeated for each year t, $t = 1970, \ldots, 2010$) and plot in Fig. 3 the coefficient on time as a measure of recent log-mortality changes (vertically) by year t (horizontally).

For example, 85-year-old male mortality (left graph, blue line) declined an average of 0.3% (i.e., a −0.003 change in log-mortality) per year between 1991 and 2000. In stark contrast, the pace of mortality decline sharply increased after 2000 (highlighted by the shaded area of the graph). Mortality for this age group declined an average of 2.6% per year between 2001 and 2010. The shaded area in the graph, corresponding to the period since the year 2000, shows that pace of mortality decline quickened for every age group for both men and women since about 2000. We will return to this key result from Fig. 3 in Section 3. These ever faster reductions in mortality may be due, in part, to greater use of statins that reduced cardiovascular disease and more widespread cancer screening. Whether or not one would regard this dramatic change as predictable ahead of time, the changes certainly became clear after a few years of unexpected declines.

2.3 *Forecasting Procedures*

SSA employs forecasting procedures that require numerous ad hoc and interrelated qualitative judgments, in a manner very difficult for any human to do well (Soneji and King 2012). For example, a critical aspect of the SSA forecast relies on the choice of 210 interrelated "ultimate rates of decline (UROD)" in mortality rates across five broad age groups, two sexes, seven causes of death, and three cost scenarios. Since 2012, the number of causes of death decreased to 5 and so the number of UROD that SSA must concurrently select became 150, which is still far too many to handle qualitatively. Rather than forecast mortality directly, SSA first assigns an annual rate of decline for each of the 75 years of its forecast. The rate of decline for the first 2 years of the forecast equals the historical rate of decline. The rate of decline for the next 23 years of the forecast linearly changes from the historical rate of decline to the subjectively chosen ultimate rate of decline. The ultimate rate of decline applies for the next 50 years of the forecast (the 26th through 75th years). SSA then imposes an additional step in its forecasts: if the historical rate of decline is negative, the rate of decline for the first 2 years of the forecast equals 75% of the historical rate of decline. Once SSA assigns rates of decline for each year in the forecast, it then iteratively multiplies the mortality rate in year t by its corresponding rate of decline. SSA then sums the forecasts across causes to produce a forecast for total mortality. Finally, SSA evaluates the quality of its overall forecast of total mortality well into the future. For example, if the age profile of the forecast in the year 2100 is not smooth or does not follow the ubiquitous shape of age profiles, SSA will readjust some of the UROD and reevaluate the quality of the updated total mortality forecast.

Formally, let $m_{a,t,c}$ represent the mortality rate for age group $a \in \mathcal{A}$, in year $t \in \mathcal{T}$, and for cause $c \in \mathcal{C}$. Let t_0 represent the year of the Trustees Report. Let $\hat{\beta}_{1,a,c}$ represent the estimated slope of a linear regression of the logarithm of historical mortality rates for age group a and cause c over the past 20 years as a function of time. Let $\gamma_{a,t_{\text{historical}},c}$ represent the historical rate of decline in mortality rate, which we estimate as $-\exp \hat{\beta}_{1,a,c}$. Let $\gamma_{a,t_{25},c}$ represent the ultimate rate of decline chosen by SSA, which applies for years t_{25}–t_{75} of the forecast. For years t_1–t_{75}, the rate of decline is determined by the following conditional equation:

$$\gamma_{a,t_i,c} = \begin{cases} \gamma_{a,t_{\text{historical}},c}, & \text{if } i \leq 2 \text{ and } \gamma_{a,t_{\text{historical}},c} > 0 \\ 0.75 \times \gamma_{a,t_{\text{historical}},c}, & \text{if } i \leq 2 \text{ and } \gamma_{a,t_{\text{historical}},c} \leq 0 \\ \gamma_{a,t_{\text{historical}},c} + \dfrac{\gamma_{a,t_{25},c} - \gamma_{a,t_{\text{historical}},c}}{23}(i-2), & 3 \leq i \leq 24 \text{ and } \gamma_{a,t_{\text{historical}},c} < \gamma_{a,t_{25},c} \\ \gamma_{a,t_{\text{historical}},c} - \dfrac{\gamma_{a,t_{\text{historical}},c} - \gamma_{a,t_{25},c}}{23}(i-2), & 3 \leq i \leq 24 \text{ and } \gamma_{a,t_{\text{historical}},c} \geq \gamma_{a,t_{25},c} \\ \gamma_{a,t_{25},c}, & i \geq 25 \end{cases}. \tag{1}$$

Finally, SSA iteratively multiplies mortality rates and rates of decline to forecast mortality rates,

$m_{a,t_i,c} = \gamma_{a,t_i,c}\, m_{a,t_{i-1},c}$, for $1 \leq i \leq 75$.

Since SSA's task of forecasting involves the choice of a large number of highly interdependent parameters consistent with one another (210 until 2011 and 150 since 2012), it is inevitable that human beings—trying as hard as they possibly could to be fair and complete—will miss a great deal. Humans are not equipped with the memories large enough to keep all of the interrelated choices consistent. Although many estimates of UROD will at least be plausible, although probably not optimal, some will also not make sense. In Fig. 4, we offer a few of the most problematic estimates we came across.

The top panel in this figure reports on deaths from diabetes in males and cancer in females. In both cases, the observed data (in dots on the left) indicate strong upwardly trending death rates, but the result of the parameter setting (the ultimate rate of decline) was an unnoticed and unjustified downward-sloping curve. As implausible as this change in trend is, worse still may be that the (vertical) differences in death rates between the two adjacent age groups portrayed massive and implausible changes between the observed data (left dots) and forecasts (right lines).

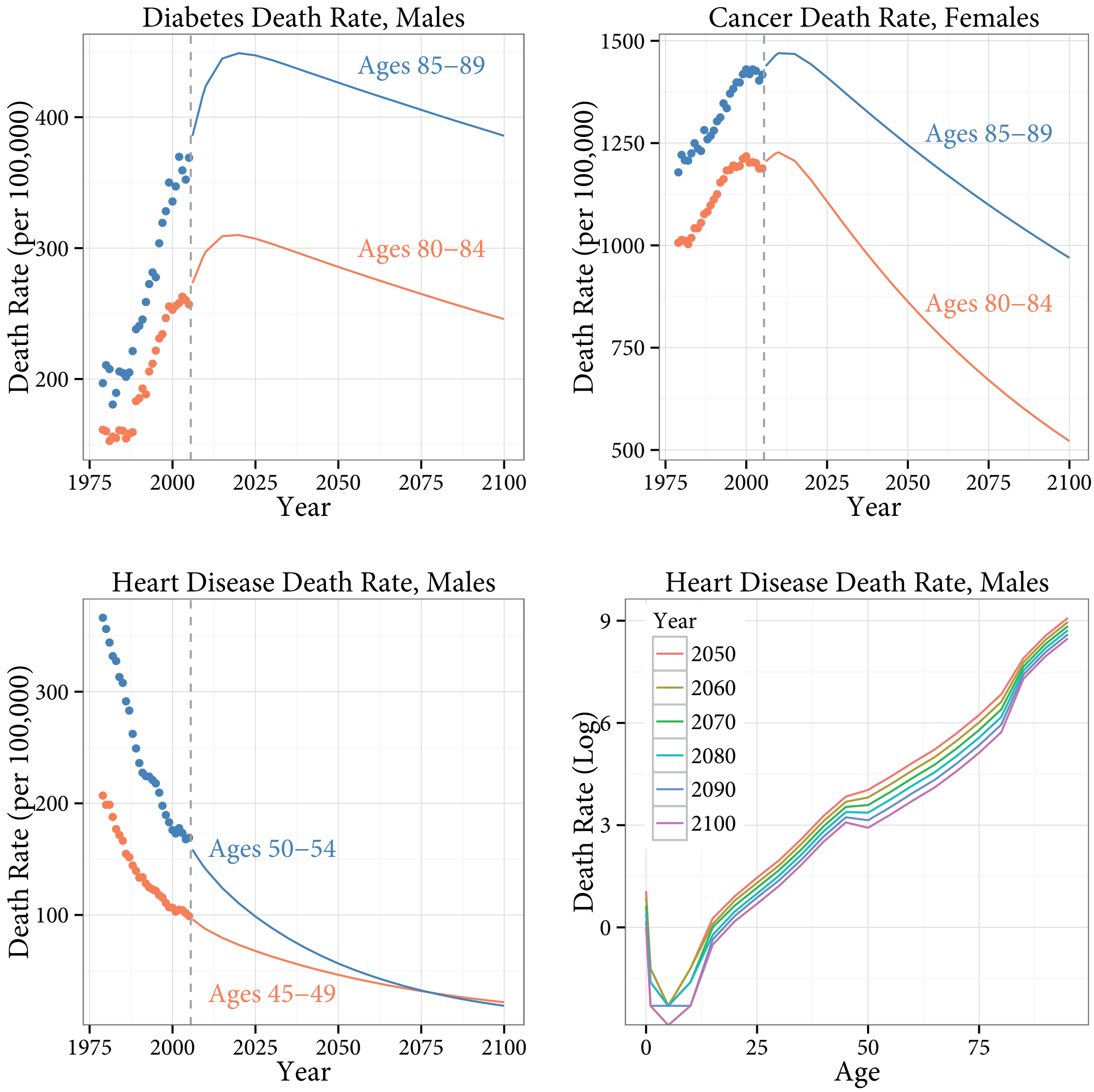

Fig. 4 Selected examples of problems with SSA's qualitative parameter selection in point estimates of cause-specific mortality forecasts.

For another example, the bottom row in Fig. 4 reports on male heart disease rates. In the left graph, the out-of-sample data show the forecasts for two age groups crossing, so that after a time it will be supposedly safer to be age 50–54 years than age 45–49 years; this, of course, makes no demographic or biological sense. Similarly, in the bottom right panel, we can see the age profile of the same data, along with a sharp drop at age 50–54 years for all the forecast years, a pattern that would only make sense if there were some strange medical discovery that did not work until age 50 years and ceased working at age 65 years. Clearly, if the experts at SSA could have focused on these numbers (as well as on all the others at the same time) or if they had encoded their knowledge in formal, reliable, and informative statistical models, these highly implausible results that are contrary to well-known demographic knowledge would not have been part of official U.S. government forecasts.

Finally, Fig. 5 gives a sample of two of the resulting SSA forecasts for male (left panel) and female (right panel) life expectancy at age 65. In this case, these are short-term (1–5-year) forecasts, but our companion paper reveals similar results for longer forecasts. The figure is presented in

Fig. 5 Residuals of shortrun forecasts of life expectancy at age 65. Solid line represents residuals based on SSA's intermediate cost scenario forecasts. The shaded region represents the interval formed by SSA's low and high cost scenario forecasts.
Note. HMD = Human Mortality Database.

terms of the residual (the forecasted SSA period life expectancy prediction minus the actual period life expectancy calculated from the Human Mortality Database; see mortality.org). In both panels, we can see approximately unbiased forecasts until about 2000 (i.e., residuals that vary around approximately zero) and then a sharp, trending, and continuing downward bias. The result after 2000 is that SSA forecasts people will die earlier than they actually do, making the Social Security Trust Funds appear healthier than they actually are.

In addition to the "best guess" intermediate cost scenario forecasts, SSA also forecasts life expectancy under low and high cost scenarios. While the interval formed by these alternative scenarios does not have a formal statistical basis, it is routinely used by the SSA to consider the range of plausible scenarios going forward. Recent Trustees Reports, for example, state that "alternatives I [low-cost] and III [high-cost] define a wide range of demographic and economic conditions." In addition to evaluating the residual for the "best guess" point estimate, Fig. 5 shows the range of residuals corresponding to the high and low cost scenarios forecast by the SSA. If this interval covers zero, it means that the truth fell within the range of the bounds forecast by SSA. Post-2000, we see that the truth always fell outside this interval.

The overall pattern in Fig. 5 is repeated in other mortality forecasts, as well as in the detailed financial forecast performance of SSA we give in Kashin, King, and Soneji (2015c). For example, the "cost rate" is one of the financial metrics the SSA forecasts and equals the ratio of the cost of the two SSA programs (Old-Age Survivors Insurance and Social Security Disability Insurance) to the taxable payroll for the year, expressed as a percentage. When comparing SSA's forecasts 5 years out to the eventual true observed value, we find that the average residual (defined as the difference between the forecast and the observed truth) from 1978 to 1999 was 0.10 percentage points. Post-2000, the average residual was −1.38 percentage points—a more than 10-fold increase in the magnitude of the error. We find a comparable large increase in error when examining forecasts 1 year out to the eventual true observed value: 0.05 average residual for forecasts made in 1978–99 and −0.52 average residual for forecasts made in 2000–11. Similar results of approximate unbiased forecasts before 2000 and very substantially biased forecasts after 2000 exist for other SSA financial forecasts we studied, including the Trust Fund balance and the Trust Fund ratio. We present detailed figures with complete evaluations of SSA forecasting errors in our companion paper

(Kashin, King, and Soneji 2015c) and its associated replication data set (Kashin, King, and Soneji 2015b).

2.4 *The Uncertainty of Policy Scoring*

In addition to producing the annual Trustees Reports, OCACT plays a singular role in American politics of scoring policy proposals put forth by members of Congress and public policy organizations. The scores are estimates of the impact of these proposals on the budget and the future of the Social Security Trust Funds. Because of the nontransparency of SSA forecasts, no other person or organization is able to produce fully independent estimates, and so all parties rely on these estimates in every major policy debate. Unfortunately, OCACT does not give uncertainty intervals for their point estimates. We fix this serious oversight here.

Since 1993, SSA has scored 105 policy proposals, including every major proposal to change the system by members of Congress and the White House. To provide rigorous uncertainty estimates, we note first that the uncertainty of counterfactual predictions, such as SSA's policy scores, always equals the sum of (a) the uncertainty of factual predictions—that is, forecasts under the current system—and (b) uncertainty due to what would happen if we change the system. Our results here quantify (a) directly and use these results as a lower bound on the total uncertainty of the policy scores.

OCACT gives estimates of the effect of policy proposals on the Trust Fund balance and on the cost rate. To estimate (a) for each, we take all forecast errors between 1 and 10 years out between 2000 and 2010, and compute the percentile of error at which each policy score appears. Policy scores that are statistically larger than zero should be larger than the 95th percentile of the forecast errors (i.e., corresponding to an $\alpha = 0.05$ significance level).

We plot all the forecast errors in Fig. 6 for the Trust Fund balance (left) and cost rate (right). For each, we have policy scores for what will happen 10 years out (on the left of each box) and 75 years out (on the right). The uncertainty estimates we present here are lower bounds; however, because they are based on forecast errors 1–10 years out, they are much less tight than lower bounds for the 75-year forecasts. Each dot represents one policy score; dots above the dashed line, appearing in red, are overwhelmed by forecasting error. The few policy scores that are significant at the 5% level appear in green at the bottom of the figure.

To be more specific, for the 75 year forecasts, the 95th percentile of forecast uncertainty for OCACT's policy scores is less than the estimated effect size of just 6 of the 85 proposal scores for Trust Fund balance and just 4 of the 25 proposal scores for the cost rate. Put differently, fewer than 10% of OCACT's 75-year-out policy scores made in the past two decades can be distinguished from random noise. None of the generally smaller 10-year-out policy scores can be distinguished from random noise. And this is under the most optimistic assumptions, with our figures being lower bounds as to the amount of uncertainty. If we were able to incorporate the likely larger uncertainty due to changing the system in ways that have never been observed (i.e., (b) above) or measure the uncertainty in real forecasts as far out as their policy scores are computed, the situation would be even more grim.

To be clear, the solution here is not for SSA to stop making policy scores, but rather for them to provide rigorous uncertainty estimates every time they make a prediction or counterfactual policy assessment.

3 The Origins of Social Security Forecasting Biases

We offer in this section a hypothesis about the origin of the systematic biases in SSA forecasting, which we summarized in Section 2. We begin by considering two possible but unlikely explanations, and then turn to our hypothesis. We present a variety of evidence for this hypothesis, including a powerful existing body of social-psychological research and numerous interviews. Then, in Section 4, we detail the internal and external pressures on OCACT, the weight of which is strongly consistent with the social-psychological evidence for what generates biases in situations like these, and how to fix them.

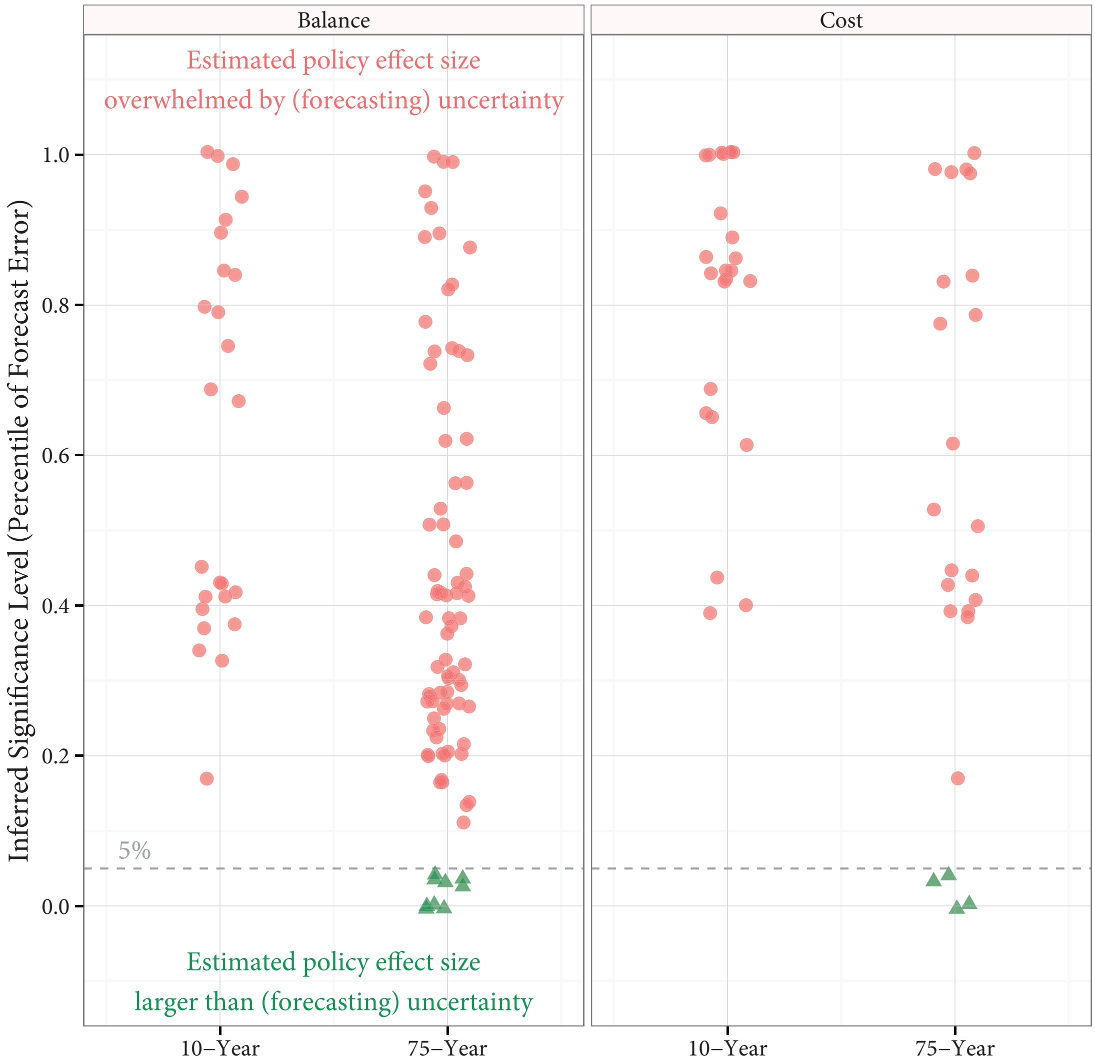

Fig. 6 Empirically inferred significance level of estimated policy effect size. Policy proposal scores for Trust Fund balance (on the left) and the cost rate (on the right), shown as red circles if forecast uncertainty overwhelms estimated policy effect size and as green triangles if forecast uncertainty (95th percentile, $\alpha = 0.05$) is less than policy effect size. At the left of each box are 10-year-out policy scores, and at the right are 75-year-out policy scores. All significance levels here are lower bounds on uncertainty, meaning that the reported α-level is likely higher than indicated.

3.1 *A Possible But Unlikely Explanation*

One logical possibility for increasing error rates in SSA forecasts is bad luck. No one predicted the onset of the Great Recession that began in December 2007, and so it would be unfair to hold SSA accountable for missing it. Additionally, between 2000 and the start of the recession, one might ask whether the small number of years of biased performance might well be due to random chance. This argument may be reasonable when evaluating the cost rate in isolation, but attributing the systematic patterns to SSA's errors in predicting life expectancy, and possibly other financial variables, is not consistent with that evidence. Moreover, SSA's errors in demographic forecasting go back significantly further than the Great Recession, and the observed values of demographic variables are highly smooth (with little change in the level of smoothness) over time. Finally, in our companion paper, we argue that the rise in unemployment that occurred during the Great Recession was not large enough to cause the corresponding decrease in mortality rates.

3.2 *A Hypothesis*

Our hypothesis for SSA's systematic forecasting errors begins with the observation portrayed in Fig. 3: mortality and the change in mortality have been trending downward since 2000. Whether this pattern was predictable or not, it became increasingly clear after a few years that mortality rates were decreasing at a faster rate and, consequently, adults age 65 years and older were living longer lives and would be drawing Social Security retirement benefits longer than anticipated. At the same time, three separate phenomena caused SSA to not respond to these dramatic changes in the input data.

First, the *possibility of bias* exists because OCACT's forecasting procedures meet essentially all the major conditions for generating inadvertent biases established in the extensive and well-documented social-psychological literature (Wilson and Brekke 1994; Gilbert 1998; Kahneman 2011; Banaji and Greenwald 2013). As we detail in Section 4.2, OCACT works hard to uphold its reputation and central position in policy debates, sometimes even to the extent of intentionally degrading the accuracy of its forecasts, by keeping them unchanged in the face of changing input data, to maintain its privileged central position. OCACT is represented at all major decision points and is open to talking to everyone in and out of government, but they see themselves—and attempt to maintain their position—as the sole judge and jury, and consequently routinely ignore recommendations of their scientific advisers. OCACT does not benefit from well-known internal procedures or external checks that could be imposed to avoid bias. OCACT relies heavily on informal decision-making procedures executed by committees composed of their staff with high levels of individual discretion and few formal procedures. In doing so, OCACT does not follow widespread objective and systematic administrative and statistical procedures that might prevent these problems. Their lack of transparency also leads to OCACT being the monopoly supplier of independent forecasts and policy proposal evaluations, with no one else in SSA, the government, or the public able to offer an alternative view or to help SSA by surfacing biases that may inadvertently arise. Although the Congressional Budget Office evaluates policy proposals, its solvency forecasting model assumes the veracity of SSA's demographic forecasts.

Second, a higher *probability of bias*, as well as the specific direction of bias, appears to have occurred because of new and unprecedented external political pressures on Social Security that began after about 2000. As we detail in Section 4.3, the Chief Actuary and his office found (or placed) themselves in the untenable position of simultaneously being the defender of Social Security, a supposedly unbiased arbiter between increasingly polarized political parties, and a defender of their own office's reputation. OCACT responded to this pressure by hunkering down and trying as hard as possible to resist change in response to political pressures. Although a laudable attempt, social psychological evidence indicates that ad hoc and qualitative procedures allow biases no matter how hard individuals try to avoid them. The particular direction of bias turned out to cause OCACT to be insulated not only from inappropriate political pressure but also from needed changes due to changing patterns in mortality and other inputs into the forecasting process.

The third and final component of our hypothesis is that we need not assume anything but *good intentions* of all employees of SSA. Our results are consistent with the oft-stated insistence of SSA officials that they try as hard as they can to be as unbiased and objective as possible with regard to external political or other pressures. Indeed, our interviewees indicate that the actuaries at OCACT, usually represented by Chief Actuary Stephen Goss, work hard to help those on both sides of every policy debate over Social Security. Some emphasized how hard the actuaries work by explaining that civil servants do not need to be at White House or Congressional policy meetings late into the night but, if that was when the discussions were happening and they could have influence, OCACT always occupied their seat at the table. OCACT jealously guards the independence it has been granted by Congress, despite pressures from members of Congress, the Administration, and other SSA officials. Moreover, OCACT does what it can to be helpful to those involved in crafting legislation while trying to avoid bias. OCACT is regarded as among the better forecasting groups in the federal government, and indeed most other countries with public retirement systems similar to SSA have no forecasting arm at all.

4 Pressures on the Social Security Administration

We now discuss social-psychological pressures, pressures internal to SSA, and external pressures on SSA.

4.1 *Social-Psychological Pressures*

Although perhaps counter intuitive, Banaji and Greenwald (2013) and numerous others in the literature have shown that good intentions can coexist with a high probability of bias when human beings perform complex tasks with high levels of discretion over many individual decisions, little feedback on whether they made the right choice the last time, high levels of external pressure, and few external checks. Humans have limited access to their own mental processes, and their biases do not give rise to any self-evident subjective experience they can use to avoid the biases (Wilson and Brekke 1994). Controlling one's own mental processes to avoid bias is often difficult or impossible, and most people vastly overestimate their ability to control their own mental processes and potential biases, even when explicitly told about documented biases in their own behavior. In fact, subject matter experts overestimate their ability to control their own personal biases even more than nonexperts.

Most importantly, attempting to reduce these biases by merely "trying harder," or replacing one person with another who is even more vigilant, will usually have little or no effect. In this regard, even "teaching psychology is mostly a waste of time" (Kahneman 2011, 170).

These psychological biases are exacerbated by problems with uncertainty estimates in the same situations. Experts who use informal qualitative approaches are typically overconfident. Indeed, the more prominent or central a role the forecaster has—and as the sole supplier of forecasts and policy evaluations, OCACT could hardly be more central—the more overconfident their statements (Tetlock 2005). The conclusion of the psychological literature on estimating confidence levels qualitatively is clear: "do not trust anyone—including yourself—to tell you how much you should trust their judgment" (Kahneman 2011, 240).

Research shows that "Biases in human reasoning are of two general types: those that result from the failure to know or apply an explicit rule of inference—the *failure of rule knowledge or application*—and those that result from *mental contamination* (cases whereby a judgment, emotion, or behavior is biased by unconscious or uncontrollable mental processes)" (Wilson and Brekke 1994, 118). A key finding of this literature is that although mental contamination in individual judgment (i.e., "personal bias") is in many instances not correctable no matter how hard one tries to avoid it, errors due to the failure of rule knowledge or application can be fixed by learning or by instituting formal procedures.

Some of the problems can be avoided altogether by replacing qualitative judgments with formal statistical rules capable of being learned and applied objectively. Others can be corrected by imposing systematic procedures on qualitative decision-making. Of course, deciding what formal statistical rules to apply are themselves also subject to mental contamination and other biases, especially when implemented by an individual or small group working together. However, these problems are vastly less likely to occur when different teams, with different backgrounds, perspectives, and preferences, check each other, as occurs in a well-functioning scientific community. This is a key advantage of following the replication standard, making data and forecasting procedures publicly available, and encouraging the scientific community and others to participate in helping ensure that SSA's forecasts are as accurate as possible.

By this theory, avoiding future forecasting biases like those we document above will probably not be achieved by OCACT personnel working harder or trying to be less biased. The solution instead is to make organizational changes that:

1. Remove human judgment where possible by formalizing informal procedures, and in the process taking advantage of dramatic progress over the past quarter century in statistical modeling and data science (about which more in Section 5);

2. Institute structural procedures when qualitative judgments are still required. For example, as in double-blind article reviews, it may be possible to elicit information from experts before they know the details of how the information they provide will affect the ultimate forecasts, and in ways that reduce the chances of "group think"; and
3. Share data and procedures with the rest of SSA, the scientific community, and the public so that any biases that inadvertently occur in steps (1) and (2) are detected and corrected, and so that groups in addition to OCACT can provide forecasts and policy evaluations.

4.2 *Internal Pressures*

The internal pressures on OCACT we detail in this section are related to the Office attempting to protect its central role in the Washington debate, to be useful to policymakers, and to be seen as important. We list here seven characteristics of OCACT that describe how it pursues these goals. These characteristics interact and overall portray hardworking public servants trying to do their jobs without any substantive bias. Yet, they remain unprotected from biases because they have not adopted the well-known procedures developed in social psychology, behavioral science, and statistics.

4.2.1 Island of fairness

First, the self-conscious public stance of OCACT is as an island of fairness and objectivity amidst a storm of partisans, and so far as we can tell this is precisely what they attempt to do. The present Chief Actuary, Stephen Goss, regularly appears in public making earnest-sounding but extreme claims about how unbiased he and OCACT are. For example, in a public address, Goss said, "I'll take a bullet before I modify anything under any kind of political pressure, and that's just an absolute. My sense is that there are some jobs and you do whatever it takes and if people don't like it that's too bad… We're giving it all we got, and objectivity, challenge everything, and no known bias, is always the mantra" (j.mp/GossCSPAN13). Similarly, one person who served as a Trustee told us emphatically, "I have never seen a *single* instance of political pressure" in OCACT. And the feeling is mutual, as indicated by Goss's public statement that the Trustees "work in a really *truly* nonpartisan way" (j.mp/GossCSPAN6-13). Having public servants who try for this level of fairness is certainly ideal but, as indicated above, trying harder to be free from bias is an ineffective way to further reduce bias unless they begin to use the well-tested advice from the scholarly literature.

4.2.2 Monopoly supplier of evaluations and forecasts

Second, as indicated in Section 1, OCACT's nontransparency, lack of data sharing, and informal forecasting methods mean it is the monopoly supplier of fully independent forecasts and evaluations of policy proposals. Goss says regularly that OCACT gives different projections under different "explicitly stated assumptions" (j.mp/GossCSPAN13); however, OCACT has full discretion to choose to evaluate or ignore any request to evaluate policy proposals under any assumptions other than those requested by Congress or the Administration. This stance may be consistent with the idea of OCACT being an island of fairness, and it may even be required, but it means that any bias inadvertently introduced, such as by choosing to evaluate only proposals with certain assumptions, is unlikely to be corrected. Moreover, SSA cannot take advantage of the central contribution of the idea of science, which is not merely "acting scientifically" but rather involves different groups checking on each other to do better than any group could on its own.

Although all final decisions are made by the Trustees, the scientific capacity to make and judge forecasts inside SSA resides almost exclusively with OCACT. Unfortunately, OCACT seems to act as a judge and jury, rather than a participant in the scientific process leading to the forecasts. Until the past 2 years, OCACT has offered little explanation as to why their forecasting practices differ so dramatically from the Technical Panels recommendations in many instances. In the 2012, 2013, and 2014 Trustees Reports, they addressed a small number of issues, mostly to declare the Technical

Panels correct or incorrect on each issue, but with little serious engagement with, or respect for, the Panel's arguments or conclusions (j.mp/OCACT12, j.mp/OCACT13, and j.mp/OCACT14).

4.2.3 Consistency bias

Third, most of our respondents emphasize that OCACT values consistency in forecasting over time above accuracy at any one time. That is, in the face of new predictive information or new methods of analysis, OCACT intentionally degrades forecast accuracy, biasing today's forecast toward yesterday's forecast. Congruent with this claim, OCACT forecasts tend to be much smoother over time than those from Medicare and others. There is of course a prudent aspect to this pattern, where a good forecaster tries not to overweight the last bit of new information.

However, many of our respondents prefer to explain this pattern via either personal or institutional explanations. The personal explanation attributes consistency bias to individuals, particularly the Chief Actuary, doing whatever he can to avoid having to admit his office was wrong. The institutional explanation attributes consistency bias to OCACT trying to emphasize its central role in the policy debate in Washington, since it recognizes that negotiation between the parties is easier when they agree on one set of consistent forecasts, even if they are wrong (or not known to be right). In support of either explanation, the bias in favor of consistency over time leads Chief Actuary Goss to be regularly described by our respondents as "intellectually biased, but not politically biased." Goss himself described this consistency bias: "the hard part is trying to balance the need to change on the basis of new ideas and understanding with the desire for consistency and stability over time" (Interview with Society of Actuaries, j.mp/GossSOA).

We can show exactly how this consistency bias occurs with an example that arose in multiple interviews we conducted. It conveys the high level of discretion allowed by OCACT's forecasting procedures (which may lead to this and other types of bias) and how consistency is implemented. When the Technical Panel strongly recommends changes in one of OCACT's myriad forecasting assumptions, they receive one of three responses: when the Chief Actuary had good evidence, he engaged the Panel and convinced them that no change was needed; when the Panel had better evidence, Goss ignored the Panel and did not change the assumptions (Autor and Duggan 2006); finally, in the small subset of cases when the Panel pushed hard even though Goss was ignoring them, Goss changed the assumption in the direction the Panel wished (although often not as far as it wished) but then changed another, unrelated, assumption not at issue in the opposite direction to counterbalance the first and keep the ultimate solvency forecasts largely unchanged.

Several of our interviewees independently suggested that they thought Goss maintains a private list of assumptions that in his best judgment require change. However, instead of making these changes when they seem to him to be scientifically warranted—immediately—OCACT introduces them over a much longer time frame, at instances specifically chosen to counterbalance other changes in the world and pressures from the Technical Panel, all in order to keep the ultimate forecasts relatively consistent over time.

The issue here, again, is not the people but the procedures, since most of these interactions are informal, out of view of the public, and thus subject to potential unintended biases. These procedures could easily be changed by making them visible, and the forecasts, as a result, would be easy to improve.

One defense of consistency bias that occasionally arises is that SSA's goal is forecasting 75 years out, and so it may not make sense to adjust forecasts in response to every new piece of information that comes in. From a Bayesian point of view, this argument is plainly false. And from the point of view of scientific evaluations, no evidence exists on the accuracy of 75-year Social Security forecasts, so all that can be done is to evaluate shorter-term forecasts where data exist. And finally, even if the shorter-term evaluations are not relevant to 75-year forecasts, they are still vitally important to tens of millions of Americans who plan to receive benefits over shorter time horizons.

4.2.4 Ignoring technical panel methodological recommendations

Fourth, SSA's external scientific advisers (their Technical Panels) have long recommended that OCACT evaluate OCACT's forecasts, share their data, make their procedures and decisions transparent, and formalize their methods. The Congressional Budget Office, for example, routinely self-evaluates its own Social Security forecasts, although they rely on OCACT demographic forecasts as an input. (The technical panel consists of outside experts appointed by the Social Security Advisory Board. The panel assesses key demographic and economic assumptions, and provides advice, but it does not independently make forecasts.) OCACT ignores or only partially follows most of the Technical Panel's recommendations. For example, on evaluating forecasts, the 2007 Technical Panel wrote:

> We believe that the accuracy of past projections should be the subject of routine reporting, either in the Trustees Report or in separate supplemental publications on methodological developments. There should be an analysis of the accuracy of past 10-, 20-, and 30-year projections similar to those periodically done by the Census, Bureau of Labor Statistics (BLS), and the Congressional Budget Office. The report should include a comparison of historical values with projected high-cost and low-cost scenario variants, noting how often each variable exceeded past projected outer bounds (j.mp/SSATech07, 4–5).

Other technical panels have also encouraged OCACT to share data and information with "different parts of SSA and within the larger research community" (j.mp/SSATech11, 3). OCACT has made some information available, and the Technical Panel has been appropriately generous in complementing OCACT for some progress, but withholding even one link in the forecasting chain means that replication is impossible. The 2007 Technical Panel was unambiguous on this point:

> Throughout this report we call for more transparency in the models and data the actuaries use, as well as the assumptions that drive their results. This recommendation is perhaps the most important one we make. Only with more transparency can other social scientists ... bring their intellect to bear on the many complex questions the Trustees and actuaries face.
>
> It is worth noting that all analytical agencies ... must make assumptions about behavior or phenomena that are unobservable or immeasurable. In the process, the assumptions become deeply embedded and the models closely guarded; and these agencies are understandably reluctant to revisit their assumptions or reveal methods. Nonetheless, it is essential to do so. First, accountability requires it ... Second, and more important, ongoing comprehensive, external review can greatly assist the quality of the analytical exercise.
>
> Transparency of the OCACT models will require several developments: (1) providing more comprehensive documentation, (2) making data from SSA records more available, (3) creating explicit models where none exist, and (4) clearly explaining the processes employed. The ultimate test of transparency is whether the actuaries' results can be essentially replicated.

These issues have been raised in many ways in all the Technical Panel reports over the past 15 years. In 2011, SSA released some appendices with additional details, but they still do not meet the replication requirements of the recommendations.

4.2.5 Ignoring Technical Panel substantive recommendations

Fifth, OCACT and the Trustees ignore, or at best undervalue and underutilize, important substantive recommendations by its Technical Panels, even those issued repeatedly. Others are given little more than token mentions or dismissals in the annual Trustees Report. On some points, the Trustees Reports directly contradict the conclusions of the Technical Panel and, in defense, the Trustees merely repeat identically worded assertions year after year in the annual Trustee Report without engaging the Technical Panel on the crucial issues raised. The Trustees and Technical Panel agree on many issues too, but the lack of engagement or mutual understanding is obvious.

Consider one important example highlighting both the lack of responsiveness of the Trustees to the Technical Panel and the high levels of discretion OCACT's actuaries have in making numerous informal and qualitative decisions—factors that create exactly the conditions for inadvertent bias more likely to occur. Part of OCACT's informal forecasting approach involves a committee choosing the large number of UROD in mortality discussed in Section 2. The committee then enters into complex discussions with many others in Congress, the Administration, and elsewhere before making a final decision. (See the Appendix for more details.)

Since human beings are incapable of keeping so many moving parts in their heads at once, the results are suboptimal and are sometimes even logically inconsistent (Soneji and King 2012). The Technical Panels have repeatedly recommended changing this approach, with little response from SSA (e.g., in 2012, the number of causes of death was reduced from 7 to 5):

> For the intermediate scenario, the Trustees make assumptions about 70 rates of decline (5 age groups × 2 sexes × 7 cause categories). Consideration of the low- and high-cost projections increases the total number of parameters to 210. The Trustees Report does not describe the process for arriving at this large number of assumptions (Technical Panel 2011, 57; j.mp/SSATech11).

> The process of producing 70 assumptions about ultimate rates of decline by age, sex, and cause of death for each of three cost scenarios is not documented and appears to be informal. Simplifying the mortality assumption will considerably improve the transparency of the Trustees Report [for emphasis, they also quote previous reports] (Technical Panel 2011, 59).

> A model based on separate projections by cause of death over a long time horizon is both implausible and inconsistent with historical experience. There is little written explanation of how these assumptions were developed (Technical Panel 2011, 59, quoting the 2003 Technical Panel, 38; j.mp/SSATech03).

In this situation, each of the 210 decisions for UROD is ambiguous; each decision depends on others and often in complicated ways. A clear definition of the objectively correct outcome does not exist, and during this process individuals in Congress, the Administration, and OCACT have their own opinions, often pushing hard to get their favored outcome. The odds are high that an individual or committee in OCACT with an even very slightly favored outcome will inevitably, perhaps imperceptibly, bias the average decision in their favor. And the evidence indicates that hardworking, professional, and objective public servants will not be able to overcome this bias, no matter how hard they try, without instituting some type of known, open, and formal procedures.

The differences of scientific opinion between the scientists on the Technical Panels and the OCACT acting on behalf of the Trustees are often of considerable importance. For example, Fig. 7 documents the four Technical Panel recommendations on the overall ultimate rate of decline of mortality (in red) and the choices made by OCACT and accepted by the Trustees (in turquoise).

The data from the four Technical Panels in this figure cover the period during which we documented systematic bias in SSA's forecasts above. The results in this example demonstrate that the Trustees have consistently ignored the Technical Panel for the past 15 years. The differences in the figure are large, consistent, and all in the same direction. In which direction has SSA chosen to deviate from its own scientific advisers for so many years? In the same direction as the systematic bias shown above in shorter-term forecasts: SSA chooses to make the Trust Funds look financially healthier than their scientific advisers think they are.

4.2.6 Informal procedures that increase vulnerability to bias

Sixth, many of OCACT's procedures have not changed in decades, despite scholarship that clearly demonstrates their suboptimality and likelihood of generating bias. Some of their procedures can be automated to eliminate bias; others cannot be automated and must remain qualitative, but could easily be changed to reduce bias.

For example, consider OCACT's informal procedures for choosing the 210 UROD (or 150 as of a few years ago). OCACT allows participants in the process to see the effect of their judgments about the value of input assumptions on the ultimate solvency forecasts while making the judgment. This means that the information they are trying to elicit is very likely to be contaminated by OCACT's known consistency bias in solvency projections (see the third point above) and possibly also any other pressure or preference.

Indeed, allowing this type of contamination is a qualitative example of a common quantitative forecasting mistake (Girosi and King 2008, 11–12), where a forecaster tweaks a statistical model until the results make the ultimate forecast consistent with their prior belief. The result, of course, is

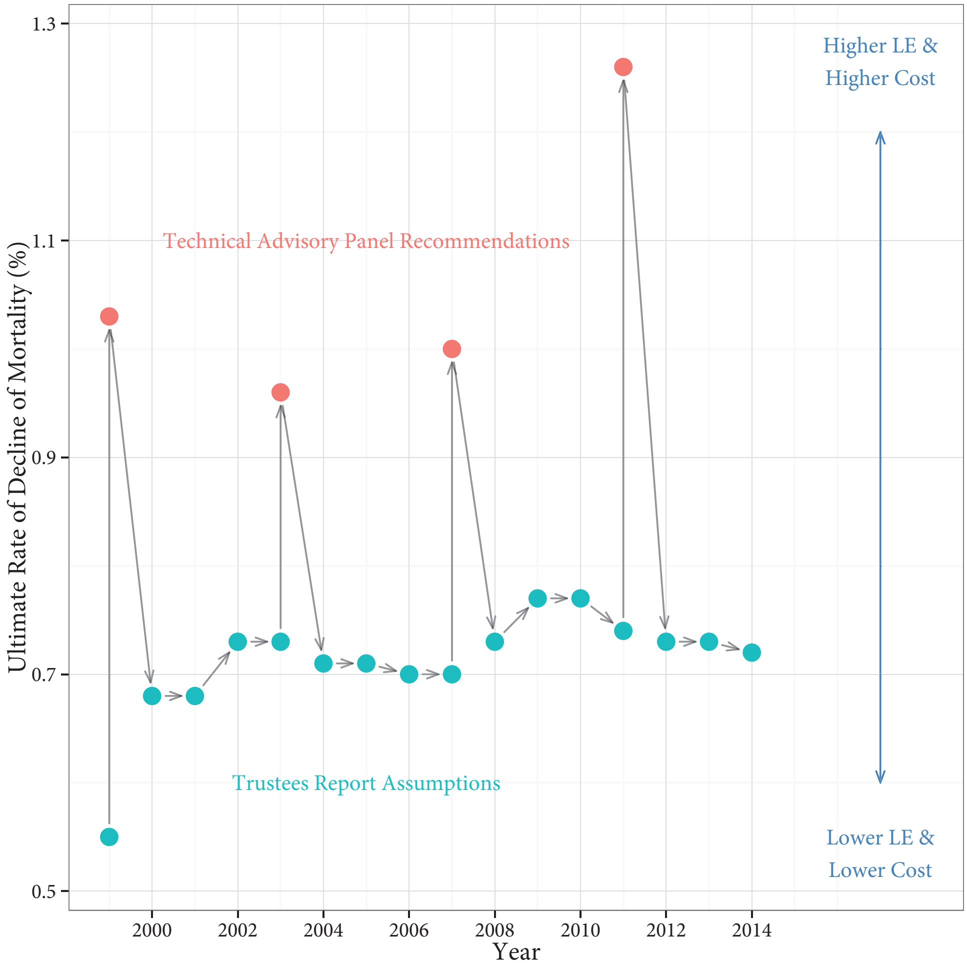

Fig. 7 Technical Panel recommendations of ultimate rate of mortality decline and SSA's chosen assumptions. Note: LE = life expectancy.

that the model itself contributes little or nothing to the process, and the forecaster is merely choosing the forecast he or she wanted initially.

Fortunately, eliminating this contamination would be straightforward and could occur in one of two ways. First, OCACT could retain their qualitative decision process and institute standard "debiasing" procedures such as blinding those providing information about the assumptions from their effects on solvency until after providing their information. Thus, OCACT would ensure that the process of eliciting information would be focused on only the information. Second, OCACT could instead formalize all available knowledge about their assumptions in a formal statistical methodology. The former would have the advantage of disrupting OCACT's practices as little as possible. The latter would change the process more, but would likely also substantially improve the quality of forecasts even after the effect of removing the bias.

4.2.7 Personnel changes

Finally, we note the most recent change in personnel in OCACT leadership: Harry Ballantyne was the Chief Actuary from 1982 to 2000; Stephen Goss took over in 2001 and remains in the post. However, attributing the increasing biases that began coincident with this change in leadership is not as obvious as may seem, as Goss had worked as a civil servant in OCACT for 30 years prior to being promoted to Chief Actuary, as have other actuaries in OCACT.

Moreover, socialpsychological evidence suggests that a change in procedure would a difference in reducing bias, unlike specific individuals following the same procedure or trying harder.

And in fact, the possibilities for bias were evident even before Goss took the helm. For example, Rosenblatt and DeWitt (2005, 44) give an example of one type of influence from an earlier era:

> "The budget assumptions of any administration are often overly optimistic," said former chief actuary Harry Ballantyne. The political leaders, whether Democrat or Republican, believe that their economic and fiscal policies will produce positive results in the short term. They would like the Trustees' Report to reflect their optimism. "Many times there are some differences" between the actuaries and the political staff members, said Ballantyne. "The cabinet staff people would say, 'why can't we assume this?' We go back and forth, and sometimes changes are made."

When many such opportunities for influence, individual discretion, and lack of formality occur, the forecasting process becomes open to the possibility of inadvertent biases that no amount of individual effort is likely to solve. Instead, the solution here is for the personnel in charge to institute some of the proven procedures well known to avoid bias. Our hypothesis is that the internal pressures discussed in this section open up the possibility for bias, with probabilities for bias increasing as a result of the increase in external pressures we discuss in the next section.

4.3 *External Pressures*

The internal situation at OCACT within SSA described in Section 4.2 makes clear that inadvertent biases are at least a possibility. Such biases can occur because numerous small informal decisions are made by people without the benefit of formal statistical models, automated computer assistance, or formal procedures for how to impartially elicit informal or qualitative information. Since many of OCACT's general procedures have remained largely unchanged for so long (during which dramatic improvements in forecasting methodology, social psychology, and behavioral science have taken place), something else must have changed to increase the likelihood of bias beyond a theoretical possibility. The hypothesis we offer here is that of unusually intense political pressures from the outside, which began at about the same time, which caused OCACT to resist all types of change, even when it was needed due to changing inputs.

To begin, we first lay out the partisan motives and political strategies underlying the ongoing debate over Social Security. In particular, the complicated politics seems to have played out in this domain in at least three ways. Through all of this OCACT was pushed and pulled from every direction, during which the actuaries tried hard to resist change.

First, the standoff on Social Security has many conservatives preferring to change the system by creating personal savings accounts managed by the government (often called "privatization," especially by opponents) or cut back its benefits, and many liberals preferring to leave the program as is or to expand it. The conservatives' political strategy to achieve these goals seems to be to emphasize the forecast point estimates, most of which predict Trust Fund insolvency by the 2030s. They then characterize the situation as a crisis and call for early negotiations to "save the program," which would likely lead to their goal of cutting it back to some degree. In contrast, the liberals' political strategy seems to be to deny any crisis by emphasizing the size of the uncertainty intervals, which are wide enough so that insolvency might well be averted by waiting for the economy to improve. Waiting might also benefit the liberals since they would have a chance for the political tide to turn and possibly emerge with a more amenable or less powerful negotiating partner (as when the original Social Security Act was passed in 1935). Of course, this characterization does not account for all liberals and conservatives. In fact, both presidential candidates in the 2012 election took what we are describing as the "liberal" position—Obama presumably for the liberal reason above and Romney to focus voters on other issues.

Second, if the two sides in this debate took the point estimates seriously (i.e., the best guess about what will happen), they might well switch positions. That is, any reasonable forecast point estimate has the system going insolvent by about the 2030s. Instead of meaning the end of the program, insolvency would translate into benefits being cut by about 25%, which might be roughly what the conservatives are after. So, waiting might be the preferred conservative strategy, whereas insisting on early negotiation—when small and relatively painless changes in payroll taxes or retirement ages could ensure long-term solvency—would be the liberal preference. Alternatively, since few involved in the political debate think it is reasonable to make changes to the system for those within a decade of retirement, waiting might well involve more taxes or fewer benefits, and so perhaps the conservatives should not want to wait. Overall, the intensity and complexity of politics here may explain some of the positions of Republicans and Democrats who, respectively, often do not fit the current conservative and liberal debating positions, and sometimes do not agree among themselves.

And finally, we could look at the issue from the perspective of time. Politicians who are trying to maximize their own self-interest need to do so within a time frame, which especially affects decision-making about a program designed to work over many generations (Weaver 1988, chap. 4). Republicans at any one point in time may want to delay shoring up Social Security in order to focus the immediate political debate on other issues, such as tax cuts. In contrast, Democrats, such as the Obama Administration taking office during an economic crisis, might also want to delay fixing long-term Social Security solvency problems while it deals with its own shorter-term issues.

The result of these highly visible conflicting political forces has led to chaotic politics over the Social Security program. However, along with the general increasing partisan rancor in Washington around the turn of the millennium, the existing political divide over Social Security began to more cleanly separate and intensify. In the late 1990s, "the politics of Social Security in the United States entered a new phase as...privatization...came to dominate the debate" (Beland 2005, 165). As opposed to previous movements to reform Social Security, privatization requires a fundamental change in the structure of the program, eliminating or reducing the importance of the Trust Fund—and at least some of the need for actuaries and OCACT to forecast its solvency—by instituting individual retirement accounts. As the conservatives increased their support for personal savings accounts, a debate about them emerged within the Clinton Administration. Although Social Security reform of some type was on their agenda, it was ultimately dropped because the Monica Lewinsky scandal used up too much political capital and attention. President Clinton, not wanting to divide his own supporters at this politically sensitive time, used his 1999 State of the Union address to oppose privatization (Beland 2005; Beland and Waddan 2012).

Around the same time, some began to argue that Social Security forecasts were too conservative, indicating that the Trust Funds were in less danger than claimed. This resonated with the Democrats, who even convinced the General Accounting Office to hire PricewaterhouseCoopers to investigate whether the "Social Security Board of Trustees may overstate the Social Security deficit" (j.mp/pwcgao). PwC evaluated whether the actuarial methods used by OCACT to produce forecasts for the 1999 Trustees Report conformed to standards of actuarial practice, such as data sources and how measures were constructed; PwC did not evaluate or compare SSA forecasts with the eventual truth.

Then, in 2001, George W. Bush took office and, beginning with his inaugural address, became the first president since the inception of Social Security to call openly for a major structural reform in the program. Bush made numerous speeches supporting personal savings accounts (j.mp/GWBss). In 2001, Bush appointed the President's Commission to Strengthen Social Security—a body instructed to support his conclusions. The Commission developed and evaluated solvency proposals under a set of guiding principles provided by the Bush Administration. While the September 11 terrorist attacks largely put Bush's social policy agenda on hold, its savings accounts were resurrected in 2005 in even stronger form (Altman 2005; Beland 2005; Beland and Waddan 2012). As part of the initiative, Bush embarked on a series of town hall meetings that represented "perhaps the most extensive public relations campaign in the history of the presidency on behalf of reforming Social Security" (Edwards 2007, 284). Sounding a favorite theme, President Bush said in a radio address that "the system is broken, and promises are being made that Social Security cannot keep" (15 January 2005, j.mp/BushRadio05). Overall, the Bush initiative "was the biggest and most concerted effort to overturn the program since its birth seventy years earlier and the only one directed from the White House itself" (Laursen 2012, 504).

These changes in the public debate put SSA and its actuaries under a new and unusually extreme form of political pressure. They were pulled in every direction—while always still trying to remain internally consistent and relatively constant over time (as described in Section 4.2). The Democratic minority on the Committee on Governmental Reform issued a partisan but compelling report with specific examples of changes in the language used by SSA as it supposedly bent to the will of the Bush White House (j.mp/DemsSS05). Contemporaneous news reports confirmed the Bush Administration's at least partially successful attempts to influence SSA communications to emphasize the financial unsustainability of the program (j.mp/NYTs1-05). In response to the claim that SSA forecasts were too conservative (and thus helping Bush's reform efforts), Charles Blahous (who was soon to become a Public Trustee of Social Security) distributed an extensive rebuttal for a public presentation (Blahous 2007, 2010). Additionally, Chief Actuary Goss openly clashed with the Republican SSA Commissioner, and Goss was, in turn, strongly defended by Democratic lawmakers (see j.mp/GossVSreps).

The external political pressure on SSA summarized here became unprecedented at just about the time when SSA's systematic forecasting biases began to increase. It appears that OCACT reacted to the extraordinarily intense and chaotic partisan politics by redoubling the practices we discussed in Section 4.2 and trying to resist the political forces, but also inadvertently resisting genuine changes in input data. They engaged more in policy discussions to try to protect their position, resisted more than ever advice from their Technical Panel and others, maintained their consistency of forecasts over time even when evidence was building to the contrary, and generally remained in a pre-2000 world.

As the data indicated change and an increasing chorus of outside voices lobbied for it, OCACT felt that the best way to retain their independence was to stay the course. Then, with the internal pressures and lack of open, transparent information and data sharing highlighted in Section 4.2, any biases that entered into the forecasting process, such as by ignoring the input data, were not likely to be detected or corrected.

The solution to these problems is not to find the perfectly unbiased and objective arbiter, since this person exists only in theory, but to open up hidden procedures for all to see and to encourage public and scientific scrutiny to help the system and its resulting forecasts improve. SSA forecasting methodology and procedures need to be modernized.

5 Formalizing Ad Hoc, Qualitative Forecasts

We provide an overview of one way to fix the problems with SSA forecasting in Section 4.1. A key component of these suggestions is to replace ad hoc qualitative forecasting with formal statistical methods that include the key features of mortality data, and so we offer one such formal statistical approach here.

We conceptualize log mortality for each time period and age group as the dependent variable in a linear regression containing time trends and risk factors such as smoking and obesity as predictors. We can then express expert information on the smoothness of log mortality via a set of specialized Bayesian priors (Girosi and King 2008). Unlike traditional Bayesian approaches, these priors are stated in terms of the expected value of the dependent variable, not in terms of coefficients. Such an approach reflects the knowledge demographers and public health experts actually have (since mortality is directly observable). In practice, placing priors on coefficients puts impossible constraints on the parameter space; priors on expected values imply priors on the coefficients that vary over the observations and, consequently, are much easier for experts to encode their prior information.

To begin, let a be the index age (for a total of A ages) and t be the index year (for a total of T years). Suppose that the log conditional probability of death at a given age a and a given year t (denoted q_{at}) follows a normal distribution, where the mean of the distribution systematically varies as a function of covariates:

$$q_{at} \sim \mathcal{N}(\mu_{at}, \sigma_a^2), \tag{2}$$

and

$$\mu_{at} = \mathbf{Z}_{a,t}\boldsymbol{\beta}_{\mathrm{a}}, \tag{3}$$

where $a = 1, \ldots, A,\ t = 1, \ldots, T$. $\mathbf{Z}_{a,t}$ is a vector of exogenous covariates that typically contains a linear time trend, smoking prevalence lagged k years in both time and age, and obesity prevalence lagged k years in both time and age. Other covariates can be easily included as well. Time is included as a crude measure of technological change; improving this measure would be valuable, and measures exist, but more predictive measures have not been found. In more general matrix notation, $\mu = \mathbf{Z}\boldsymbol{\beta}$, where $\mathbf{Z}$ is a block diagonal matrix (with $\mathbf{Z}_a$ forming the blocks) and β is a column vector formed by concatenating the A age-specific coefficient vectors.

As discussed, experts possess key information about the behavior that mortality forecasts should exhibit. This expert information is grounded in decades of carefully examining demographic data and the patterns within it. The typical Bayesian approach is to quantify such expert information in the specification of a prior on the regression, $\mathcal{P}(\beta|\theta)$, where θ is a hyperparameter or a smoothing parameter. Similarly, $\mathcal{P}(\sigma)$ is the prior distribution for the standard deviation.

We allow users to formulate their priors in terms of the expected value of log mortality (the dependent variable), subsequently backing out the corresponding prior on the regression coefficients: $\mathcal{P}(\mu|\theta) \Rightarrow \mathcal{P}(\beta|\theta)$, which then allows standard Bayesian computational techniques to be used. Since μ_i is a scalar and β is a vector, the many-to-one transformation seems impossible. However, if we restrict our attention to the subspace of $\mathbb{R}^{T\times A}$ where μ can be explained by the covariates $\mathbf{Z}$, $\mathbb{S}_{\mathbf{Z}} \subset \mathbb{R}^{T\times A}$ (which is the support of the prior), the transformation is directly invertible without additional assumptions beyond that $\mathbf{Z}$ is of full rank: $\beta = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mu$.

Then we only need to specify priors in a manner that accurately represents expert knowledge about the behavior of μ, the expectation of log mortality. We do this in a set of L statements. Each statement, written as $H_l[\mu]$ for $l \in [1, ., L]$, is a *smoothness functional* (a map from a set of functions to the set of real numbers). The prior distribution on the expected value of log mortality may then be expressed in terms of these L smoothness functionals:

$$\mathcal{P}(\mu|\theta) \propto \exp\left(-\frac{1}{2}\sum_{l} H_l[\mu, \theta]\right), \tag{4}$$

where θ_l is a non-negative weight placed on the lth smoothness functional. In practice, much of expert knowledge about the behavior of log mortality may be translated into at least three types of smoothness functionals. First, smoothness over age stipulates that adjacent age groups should have similar expected values of log mortality. Second, smoothness over time stipulates that nearby time periods should have similar expected values of log mortality. Third, smoothness over age and time stipulates that adjacent age groups should have similar time trends of the expected values of log mortality.

For example, consider smoothness over age groups. Intuitively, we want adjacent age groups to have similar expected values of log mortality, μ. How can we quantify smoothness? Many simple encodings of smoothness turn out to be implausible; for example, using the straightforward squared differences in adjacent age groups as the core of a smoothness functional would imply a random walk over age. A better option is to calculate the squared second derivative of expected log mortality:

$$\left(\frac{\partial^2\mu(a,t)}{\partial a^2}\right)^2. \tag{5}$$

The second derivative measures the curvature of $\mu(a, t)$ with respect to age. If $\mu(a, t)$ is a linear function with respect to age, then the curvature will be 0. Any deviation from a linear function has a larger squared second derivative, and thus is inherently less smooth. However, as we saw in Fig. 2, certain non-smoothness is inherent in the observed data. For example, log mortality in age group 0–4 is usually very different from log mortality for ages 5–9. Similarly, the mortality bump in late adolescence and early adulthood is a ubiquitous non-smooth feature of the age profile. To better account for these empirical regularities in demographic data, it makes sense to measure smoothness in terms of deviations from a "typical" age profile:

$$\left(\frac{\partial^2 \mu(a,t) - \overline{\mu}(a)}{\partial a^2}\right)^2, \tag{6}$$

where $\overline{\mu}(a)$ is a typical age profile, such as calculated from the average over real data. (Using the average enables us to profess ignorance over the level of the data, formally using an improper prior, putting the mean in the null space, about which more below.) Now any linear deviation from the typical age profile is considered smooth. The greater the curvature of the deviation from the typical age profile, the less smooth the behavior of the dependent variable.

To obtain a single value of smoothness across all the ages and times in a forecast, we take the expected value of the squared second derivative with respect to age and time:

$$H_{\text{age}}[\mu,\theta] = \theta_{\text{age}} \int_0^T \mathrm{d}w^{\text{time}}(t) \int_0^A \mathrm{d}w^{\text{age}}(a) \left(\frac{\partial^2 \mu(a,t) - \overline{\mu}(a)}{\partial a^2}\right)^2, \tag{7}$$

where $\mathrm{d}w^{\text{time}}(t)$ and $\mathrm{d}w^{\text{age}}(t)$ represent weights that can be used to assign greater importance to smoothness in certain ages and times than others. Then, we add a positive parameter θ_{age} that controls how influential this functional is vis-à-vis other functionals and the data. Finally, we encode the expert qualitative knowledge that the expected value of log mortality in adjacent age groups should be similar by merely specifying that $H_{\text{age}}[\mu,\theta]$ should be small.

Since the second derivative is indifferent between linear functions, our priors do not impose any specific linear relationship between age and the expected value of the dependent variable. To be specific, the null space of the prior is composed of all linear functions. Following this intuition, we can then write down the smoothness functionals for all three types of smoothness:

- Smoothness over age

$$H_{\text{age}}[\mu,\theta] = \theta_{\text{age}} \int_0^T \mathrm{d}w^{\text{time}}(t) \int_0^A \mathrm{d}w^{\text{age}}(a) \left(\frac{\partial^2 \mu(a,t) - \overline{\mu}(a)}{\partial a^2}\right)^2, \tag{8}$$

- Smoothness over time

$$H_{\text{time}}[\mu,\theta] = \theta_{\text{time}} \int_0^T \mathrm{d}w^{\text{time}}(t) \int_0^A \mathrm{d}w^{\text{age}}(a) \left(\frac{\partial^2 \mu(a,t)}{\partial t^2}\right)^2, \tag{9}$$

- Smoothness over age and time

$$H_{\text{age/time}}[\mu,\theta] = \theta_{\text{age/time}} \int_0^T \mathrm{d}w^{\text{time}}(t) \int_0^A \mathrm{d}w^{\text{age}}(a) \left(\frac{\partial^3 \mu(a,t)}{\partial a \partial t^2}\right)^2. \tag{10}$$

Our complete prior is then

$$\mathcal{P}(\mu|\theta) \propto \exp\left(-\frac{1}{2}\left\{H_{\text{age}}[\mu,\theta] + H_{\text{time}}[\mu,\theta] + H_{\text{age/time}}[\mu,\theta]\right\}\right) = \exp\left(-\frac{1}{2}H[\mu,\theta_{\text{age}},\theta_{\text{time}},\theta_{\text{age/time}}]\right), \tag{11}$$

when μ lies in $\mathbb{S}_{\mathbf{Z}}$ and 0 otherwise. That is, we restrict the prior to only expected values of log mortality that may be explained using the covariates $\mathbf{Z}$. If we restrict the prior on μ to just the subspace that spans the covariates, then we are able to express β in terms of μ:

$$\beta = (Z'Z)^{-1}Z'\mu. \tag{12}$$

Consequently, the prior on the expected value of log mortality can be translated into a prior on the regression coefficients (and where we use the superscript H^{μ}, following Girosi and King (2008, 68), to emphasize that the density is derived using knowledge of μ):

$$\mathcal{P}(\beta|\theta) \propto \exp\left(-\frac{1}{2}H[Z\beta, \theta_{\text{age}}, \theta_{\text{time}}, \theta_{\text{age/time}}]\right) = \exp\left(-\frac{1}{2}H^{\mu}[\beta, \theta_{\text{age}}, \theta_{\text{time}}, \theta_{\text{age/time}}]\right). \tag{13}$$

The resulting prior is improper since, as $\theta \to \infty$, the prior collapses to a projection in the null space, not a single point. In this null space, the forecasts are entirely dependent on the likelihood and thus the data.

Inherent in this procedure is a trade-off between smoothness and predictive accuracy. This trade-off may be controlled by setting θ_{age}, θ_{time}, and $\theta_{\text{age/time}}$. For examples of forecasts from this type of model, see Girosi and King (2008), King and Soneji (2011), and Soneji and King (2012).

Model-based uncertainty estimates and credible confidence intervals for this forecasting procedure may be obtained using Gibbs sampling, whereby we sample from the posterior of β and calculate the forecast log mortality for each iteration. In order to account for model dependence—likely the greatest source of uncertainty—we can take a robust Bayesian approach by supplying a range of prior values where θ must lie (King and Zeng 2004). And in order to more formally extract information about the hyperprior parameters from experts, we could use more formal methods for Bayesian elicitation.

If OCACT employed this statistical forecasting model, rather than its ad hoc qualitative forecasting model, it could encode its expert judgment in the selection of covariates, lag specification between covariates and mortality, and the choice of Bayesian priors. In doing so, OCACT would eliminate the informal procedures that increase vulnerability to bias and produce a transparent and scientific model that may be improved upon.

6 Concluding Remarks

In this study, we offer a possible explanation for the systematic biases we found in Social Security's demographic and financial forecasts in Kashin, King, and Soneji (2015c). The possibility of bias arises because of the lack of professionally designed formal procedures in place to avoid them. Unnecessary informality and ad hoc qualitative forecasting approaches, lack of up-to-date statistical methods that could automate decisions considerably too difficult to manually make individually or collectively, and the absence of transparency (e.g., publicly available replication information) open up SSA to the possibility of bias. OCACT adds to the problems by insisting on the consistency of forecasts over time, in order to remain at the center of the policy debate, even when contradicted by strong trends in the data and SSA's Technical Panel experts.

Trying to resist the continuing intense political pressure is just what Americans would want of their government officials. Yet, the difficulty of the task and their suboptimal procedures caused the actuaries to hunker down and resist all dynamics that might lead them to modify their forecasts, including genuine changes in patterns in demography, public health, risk factors, or medical technology. These dynamics combined with the perfect storm of political pressure, suboptimal forecasting procedures open to bias, and changes in the world (e.g., faster increases in life expectancy). Thus, as SSA continued to resist modifying their forecasts, doing so led to a much higher probability of bias.

Our study also identifies steps SSA can take to reduce the biases in its forecasts. SSA would benefit from instituting formal procedures within OCACT to learn from its forecasting errors and to reduce its biases. Additionally, SSA should seriously engage the issues raised by its own scientific advisers. As the Technical Panels emphasize, open evaluation of past performance is the best way to guarantee that forecasters learn over time, which is why open, repeated evaluation is common throughout other parts of government, commerce, industry, and academia. At least from now on, every SSA Trustees Report should routinely provide a comprehensive evaluation of prior forecasts.

Furthermore, SSA should also use what it learns from each evaluation to refine subsequent forecasts. They should openly share their data and methods with the public so that members of the scientific community can easily replicate SSA's forecasts and contribute to their improvement. And, importantly, they should institute structural barriers to prevent inadvertent bias in the form of more formalized and transparent statistical procedures that are also less subject to manipulation and mental contamination. As it happens, these procedures are not only more replicable and easier to share with others; they also enable SSA to take advantage of the spectacular advances in statistics, data analytics, demography, and machine learning over the past several decades. Adopting these procedures will improve the quality of SSA's forecasts, correct the obvious biases we unearthed above, and help ensure open and fully informed democratic debate.

Fair, transparent, and accurate forecasts afford members of Congress the ability to consider alternative assumptions as they debate policy proposals to preserve the solvency of Social Security. SSA's failure to follow well-developed best scientific practices represents a significant squandering of public resources and hampers meaningful progress on policy changes to Social Security. If OCACT relinquishes its monopoly position as the sole provider of both demographic and financial forecasts and fully reveals its forecasting procedures, it will advance its own objective of producing the best and least biased forecasts possible. Moreover, this reform may also improve forecasts of the Medicare Trust Funds, which rely on SSA OCACT demographic forecasts as an input.

Humanity has not yet found a better way to learn than the collective efforts of the scientific community pursing the same goals, most of which would come as free effort to OCACT, SSA, and the U.S. government. Regardless of the causes, however, fixing these problems is not only crucial for SSA. For the future of Social Security, and even for American democracy, the alternative assumptions of those debating policy proposals must be based on fair, open, and accurate forecasts.

Funding

Samir Soneji was also supported by grant KL2TR001088 from the National Center for Advancing Translational Sciences.

A Data Appendix

Observed Life Tables

We obtain observed conditional probability of death and period life expectancy data used in Figs. 1, 3, and 2 from the Human Mortality Database (2015). Figure 1 plots conditional probabilities of death in the United States from 1980 to 2010 for males and females separately. Figure 2 plots conditional probabilities of death across 39 countries for both genders combined. The availability of the data ranges from as early as 1751 for Sweden until 2013 at the latest. Complete details of data availability by country are available at j.mp/HMD availability. Figure 3 plots the changes in conditional probabilities over successive 10-year intervals from 1960 to 2010. We present the results by age (65, 75, 85, and 95) and sex. This change is calculated as the slope of a regression of the log of the conditional probability of death on time for each 10-year period. For example, the value for 1970 is the slope in the aforementioned regression for the 10-year window from 1961 to 1970. In the plot, we weight each age in proportion to the number of deaths for that age in 2010.

SSA's Forecasts of Cause-Specific Mortality

We obtained observed (1979–2005) and forecast (2006–2100) cause-specific mortality rates under SSA's intermediate cost scenario from personal communication with Felicitie Bell on May 15, 2009.

SSA's Forecasts of Life Expectancy

We collected all life expectancy forecasts published in the annual Trustees Reports from the Board of Trustees (1982–2010). In reports prior to 2001, SSA published life expectancy at birth and at age 65 years for males and females projected in 5-year intervals for a total of 75 years into the future. Post-2001, supplementary single-year tables are included online. Our sources are Table 11 of Trustees Reports 1982–1991, Table II.D.2 of Trustees Reports 1992–2000, and Table V.A3 of the supplementary single-year tables of Trustees Reports 2001–2010. We calculate short-term residuals as the difference between SSA's "best guess" projection (intermediate-cost scenario/ alternative II) for the next year divisible by 5 and the observed life expectancy from HMD.

UROD of Mortality

UROD are single-number summaries of the myriad long-term mortality assumptions made by the Trustees by sex, age group, and cause of death. They represent the average annual percent reduction in age-adjusted central death rates for the last 50 years of the 75-year projection window. We gather the intermediate-cost scenario ("best guess") URODs assumed by 1991–2013 Trustees Reports and the URODs recommended by the four Technical Panels commissioned by the Social Security Advisory Board (quadrennially starting with 1999). For the 1999 TR, we average the male and female UROD as reported in the 2003 Technical Panel Report (Social Security Advisory Board Technical Panel 2003). Page 71 of the 2001 TR gives the UROD assumed by 2000 and 2001 TR. Table II.C1 of TR 2002–2009 gives the assumed UROD. UROD for 2010 TR is found on page 80 of the report. For 2011, we average the male and female UROD—0.75% and 0.73%, respectively—as reported in Table 2.3 of Office of the Chief Actuary (2013). For 2012, 2013, and 2014 Trustees Reports, the UROD are found in Table 2.2 of the respective *Long-Range Demographic Assumptions* (Office of the Chief Actuary 2012, 2013, 2014).

UROD recommendations made by the 1999 and 2003 Technical Panel are available in Table 3 of the 2003 Technical Panel Report (Social Security Advisory Board Technical Panel 2003). The recommended UROD by the 2007 Technical Panel is available in Table 1 of their report (Social Security Advisory Board Technical Panel 2007). Although the 2011 Technical Panel doesn't make its recommendations in terms of UROD, *The Long-Range Demographic Assumptions for the 2013 Trustees Report* converts the 2011 Technical Panel's life expectancy recommendation into an UROD.

Policy Proposal Scoring

We collected all policy proposals submitted by Congress and select public policy organizations (http://www.ssa.gov/oact/solvency/). The Social Security Administration Office of the Chief Actuary estimated the effect of each proposal on Social Security solvency (e.g., the cost rate and the Trust Fund balance). We denote the estimated effect of a policy proposal as its 'score'.

References

Altman, Nancy J. 2005. *The battle for Social Security: From FDR's vision to Bush's gamble*. Hoboken, NJ: John Wiley & Sons.

Autor, David H., and Mark G. Duggan. 2006. The growth in the Social Security disability rolls: A fiscal crisis unfolding. *Journal of Economic Perspectives* 20(3):71–96.

Banaji, Mahzarin R., and Anthony G. Greenwald. 2013. *Blindspot: Hidden biases of good people*. New York, NY: Delacorte Press.

Beland, Daniel. 2005. *Social Security: History and politics from the New Deal to the privatization debate*. Lawrence, KS: University Press of Kansas.

Beland, Daniel, and Alex Waddan. 2012. *The politics of policy change: Welfare, Medicare, and Social Security reform in the United States*. Washington, DC: Georgetown University Press.

Blahous, Charles P., III 2007. Have the Social Security Trustees been too conservative? Paper presented at the American Enterprise Institute.

———. 2010. *Social Security: The unfinished work*. Stanford, CA: Hoover Institution Press.

Edwards, George C. III. 2007. *Governing by campaigning: The politics of the Bush presidency*. New York, NY: Longman Publishing.

Gilbert, Daniel T. 1998. Ordinary psychology. In *The handbook of social psychology*, eds. Daniel T. Gilbert, Susan T. Fiske, and G. Lindzey, Vol. 2, 89–150. New York: McGraw Hill.

Girosi, Federico, and Gary King 2008. *Demographic forecasting*. Princeton, NJ: Princeton University Press. http://gking.harvard.edu/files/smooth/ (accessed April 29, 2015).

Human Mortality Database. 2015. University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). or (data downloaded on 31 January 2015) http://www.mortality.org or http://www.humanmortality.de (accessed April 29, 2015).

Kahneman, Daniel. 2011. *Thinking, fast and slow*. New York, NY: Farrar, Straus and Giroux.

Kashin, Konstantin, Gary King, and Samir Soneji 2015a. Replication data for: Explaining systematic bias and nontransparency in U.S. Social Security Administration forecasts. Harvard Dataverse [Distributor] V1 [Version]. UNF:6:967llFHgiywsHWWp1cVg9A== http://dx.doi.org/10.7910/DVN/28323 (accessed April 29, 2015).

———. 2015b. Replication data for: Systematic bias and nontransparency in U.S. Social Security Administration forecasts. Harvard Dataverse [Distributor] V1 [Version]. UNF:5:1oerGFXQ0Bu9bcMFU5/t2A== http://dx.doi.org/10.7910/DVN/28122 (accessed April 29, 2015).

———. 2015c. Systematic bias and nontransparency in U.S. Social Security Administration forecasts. *Journal of Economic Perspectives*. In press.

King, Gary. 1995. Replication, replication. *PS: Political Science and Politics* 28(3):443–99. http://j.mp/jCyfF1 (accessed April 29, 2015).

King, Gary, and Langche Zeng 2004. Inference in case–control studies. In *Encyclopedia of biopharmaceutical statistics*, ed. Shein-Chung Chow, 2nd ed. New York: Marcel Dekker. http://gking.harvard.edu/files/abs/1s-enc-abs.shtml (accessed April 29, 2015).

King, Gary, and Samir Soneji. 2011. The future of death in America. *Demographic Research* 25(1):1–38. http://j.mp/iXUpBv (accessed April 29, 2015).

Laursen, Eric. 2012. *The people's pension: The struggle to defend Social Security since Reagan*. Edinburgh, Scotland: AK Press.

Lee, Ronald D., and Lawrence R. Carter. 1992. Modeling and forecasting U.S. mortality. *Journal of the American Statistical Association* 87(419):659–75.

Office of the Chief Actuary. 2012. The long-range demographic assumptions for the 2012 Trustees Report. Technical report, Social Security Administration. http://j.mp/OCACT12 (accessed April 29, 2015).

———. 2013. The long-range demographic assumptions for the 2013 Trustees Report. Technical report, Social Security Administration. http://j.mp/OCACT13 (accessed April 29, 2015).

———. 2014. The long-range demographic assumptions for the 2014 Trustees Report. Technical report, Social Security Administration. http://j.mp/OCACT14 (accessed April 29, 2015).

Rosenblatt, Robert, and Larry DeWitt 2005. The role of Social Security's chief actuary. *Contingencies* (July/August):40–5. http://j.mp/ChiefA (accessed April 29, 2015).

Social Security Advisory Board Technical Panel. 2003. Technical Panel on assumptions and methods. Technical report, Social Security Advisory Board. http://j.mp/SSATech03 (accessed April 29, 2015).

———. 2007. Technical Panel on assumptions and methods. Technical report, Social Security Advisory Board. http://j.mp/SSATech07.

———. 2011. Technical Panel on assumptions and methods. Technical report, Social Security Advisory Board http://j.mp/SSATech11.

Soneji, Samir, and Gary King. 2012. Statistical security for Social Security. *Demography* 49(3):1037–60. http://j.mp/Qvla7N (accessed April 29, 2015).

Tetlock, Philip E. 2005. *Expert political judgment: How good is it? How can we know?* Princeton, NJ: Princeton University Press.

The Board of Trustees, Federal OASDI Trust Funds. 1982–2010. Annual report of the Board of Trustees of the federal old-age and survivors insurance and federal disability insurance Trust Funds. Technical report, Social Security Advisory Board.

Weaver, R. Kent. 1988. *Automatic government: The politics of indexation*. Brookings Institution Press.

Wilson, Timothy D., and Nancy Brekke. 1994. Mental contamination and mental correction: Unwanted influences on judgments and evaluations. *Psychological Bulletin* 116(1):117.