Review of Behavioral Economics, 2016, 3: 189–219

Consistent Bayesians Are No More Accurate Than Non-Bayesians: Economists Surveyed About PSA

Nathan Berg[1], Guido Biele[2] and Gerd Gigerenzer[3]

[1] *Department of Economics, University of Otago, New Zealand*
[2] *Division of Mental Health, Norwegian Institute of Public Health, Oslo, Norway*
[3] *Max Planck Institute for Human Development, Berlin, Germany*

ABSTRACT

This paper looks for, but cannot find, evidence that links belief inconsistency to belief inaccuracy or economic loss. Economists with consistent (i.e., Bayesian) conditional beliefs about the sensitivity and positive predictive value of the Prostate Specific Antigen (PSA) test have unconditional beliefs about the risk of prostate cancer that are, if anything, less accurate than the beliefs of inconsistent non-Bayesians. PSA decisions depend more on the advice of doctors and family members than on beliefs about cancer risks. Men's beliefs about the pros and cons of PSA testing do not explain self-reported PSA decisions. This absence of evidence that non-Bayesian beliefs lead to economic loss suggests that belief consistency may be relatively unimportant as a normative criterion in high-stakes decision tasks that reward accuracy instead of consistency. A technique is introduced for eliciting measures of both consistency and accuracy of an individual's probabilistic beliefs.

Keywords: Logical consistency, Predictive accuracy, Elicitation, Non-Bayesian, Ecological rationality, Medical decision making

JEL Codes: D03, D6, D8, A11, C11

ISSN 2326-6198; DOI 10.1561/105.00000034

> For judged probabilities to be considered adequate, or rational, internal consistency is not enough.
>
> – Tversky and Kahneman (1974)

> It appears that a minimal requirement of rationality is that one not hold beliefs that are contrary to objectively available data, coupled with logical, statistical, or mathematical reasoning.
>
> – Gilboa, Postlewaite and Schmeidler (2009)

1 Introduction

Consistency of prior and posterior beliefs (i.e., conforming to Bayes' Rule) is the predominant normative characterization of what it means to have rational beliefs.[1] Gilboa *et al.* (2010), for example, write: "The mode of reasoning most widely used in economic modeling is *Bayesian*." Starmer (2000) observes that before non-additive probability models appeared in the economics literature, economists usually took it for granted that the Savage Axioms (guaranteeing that choice over lotteries can be represented as expected utility maximization with respect to subjective belief distributions that conform to Bayes' Rule) provide the "right model of individual choice." Selten (2001) writes that "[m]odern mainstream economic theory is largely based on an unrealistic picture of human decision making [in which] agents are portrayed as fully rational Bayesian maximizers of subjective utility." Camerer *et al.*'s (2003) definition of "full rationality" requires that "people have well-formed beliefs about how uncertainty will resolve itself, and when new information becomes available, they update their beliefs using Bayes's law." According to Aragones *et al.* (2005), "[m]ost of the formal literature in economic theory and in related fields is based on the Bayesian model of information processing." And Gilboa *et al.* (2009) emphasize the singularity of Bayesian information processing (as opposed to a plural toolkit of mechanisms that could be used to formulate reasonable beliefs), stating that "within economic theory the Bayesian approach is the sole claimant to the throne of rationality."[2]

Despite the normative force of internal logical consistency that characterizes Bayesian beliefs, a distinct (and in some cases perhaps more compelling)

[1] Savage (1954) argued for a normative interpretation of expected utility theory while admitting to violating its consistency requirements when first encountering Allais' paradox. See Starmer (2000, 2009) for more on normative interpretations of expected utility theory.

[2] Gintis (forthcoming) states strong support for Bayesian consistency as a universal assumption: "I have always been comfortable with identifying rationality with the Savage axioms, which may be described in shorthand as 'preference consistency over lotteries with subjective probabilities."' Loewenstein (2006) urges caution, however. Cubitt and Sugden (2001) show that inconsistent individuals do not always succumb to exploitative competitors.

normative criterion for evaluating subjective beliefs is accuracy. There is no mathematical or analytic requirement that Bayesian beliefs are any more accurate (with respect to objective frequency distributions) than non-Bayesian beliefs. Conditional beliefs can be perfectly consistent yet grossly inaccurate. Therefore, it is an empirical question as to whether Bayesian beliefs tend to be any more (or less) accurate. Surprisingly, there is, as yet, little empirical evidence associating logical consistency to objective accuracy. To pursue this empirical question, this paper reports data collected from economists addressing the following three objectives.

(i) We look for evidence that inconsistency (i.e., violations of the normative criterion of conforming to Bayes' Rule) affects the expected inaccuracy of subjective beliefs (violations of the normative criterion that beliefs are closely calibrated to objective frequencies). Unconditionally and conditionally, we find no positive statistical associations between these two distinct normative criteria.

(ii) We test whether inconsistency affects choices over actions, in this case whether beliefs about the PSA test and risks of prostate cancer have an effect on the probability that a man over 40 chooses to have a PSA test. Unconditionally and conditionally, we find no evidence that inconsistency influences PSA testing decisions.

(iii) We test whether subjective beliefs about the risks of prostate cancer and PSA testing, including possible harms, jointly affect the probability of PSA testing. We find no evidence to reject the null hypothesis that PSA decisions are independent of beliefs about both disease frequency and intensity of harm until controls for social influences are included in the empirical model.

This paper introduces a technique for eliciting information about both the consistency and accuracy of an individual's beliefs. Inconsistency is measured by comparing the ratio of conditional beliefs to the ratio of unconditional probabilities we provided. Inaccuracy is measured by comparing unconditional beliefs to published point estimates of those unconditional probabilities. To our knowledge, the belief data we report provide the first empirical test of whether people with logically inconsistent (i.e., non-Bayesian) beliefs are any less accurate. Caution is, of course, warranted when interpreting absence of evidence that inconsistency and inaccuracy are unconditionally or conditionally correlated (i.e. failing to reject a null hypothesis of zero correlation).[3] To the

[3]The absence of correlation between inconsistency and inaccuracy reported in this paper is not easily dismissed as the result of low statistical power. Given our sample size, testing the null hypothesis that the Pearson correlation coefficient is zero when the true correlation is 1/3 (using Fisher's transformation to compute the power function for a two-sided test)

extent that this absence of evidence linking consistency and accuracy of beliefs is real, the normative force of Bayes' Rule in settings where accuracy rather than consistency is rewarded may be called into question.[4]

A second, more challenging issue is whether inconsistency is associated with economic losses. Despite the vast literature on non-Bayesian beliefs, one finds surprisingly little evidence to substantiate the hypothesis that deviations from Bayes' Rule generate meaningful losses.[5] Raising questions about whether deviations from orthodox requirements of rationality based on internal consistency such as Bayes' Rule are costly (or perhaps beneficial) should not imply broader skepticism about the experimental evidence documenting those anomalies and biases. On the contrary, when one takes the behavioral economics literature seriously, especially its priority on empirical realism, it suggests a much needed follow-up question: If individuals do not conform to norms of internal logical consistency, what then is the economic cost?

Section 2 describes the data. Section 3 reports evidence linking belief consistency to PSA decisions. Section 4 concludes with interpretations of the empirical results.

2 Description of Data

2.1 Summary Statistics of Survey Respondents

We surveyed attendees of the annual meeting of the American Economic Association (attended by more than 10,000 registered conference participants), also known as the Allied Social Science Associations. Our interviewer conducted

gives a chance of rejecting the null of 96.7%. The no-decision classification (NDC) procedure for measuring strength of evidence while allowing for three decision outcomes (Berg, 2004) similarly points toward rejecting reasonably sized alternative hypotheses rather than the evidentially ambiguous "no decision". NDC partitions the space of the test statistic into three regions corresponding to (a) reject the null at a specified probability of type-1 error, (b) make no decision, or (c) reject the alternative hypothesis at a specified probability of type-2 error.

[4]Psychologists Hastie and Rasinski (1988) appear to be the first to have classified normative criteria according to whether these require internal consistency or what they refer to as *correspondence* (sufficiently high performance by a free-standing metric such as objective accuracy). Hastie and Rasinski (1988) and Hammond (1996) refer to norms based on internal consistency as *coherence norms*, which include Bayes' Rule, the Kolmogorov axioms, and transitive preferences. In contrast, *correspondence norms* evaluate beliefs and decisions by how well they correspond to the demands of the decision maker's environment (not based on internal consistency, e.g., accuracy of beliefs, accumulated wealth, lifespan, or happiness). Coherence norms impose restrictions on pairs or larger sets of beliefs or decisions belonging to a single decision maker. In contrast, correspondence norms enable interpersonal rankings on the basis of a single observation from each decision maker. Gilboa (2010) argues in favor of considering normative criteria other than consistency.

[5]Behavioral economists have paid close attention to modeling and empirically documenting deviations from Bayes' Rule (e.g., Camerer, 1987; Ganguly *et al.*, 2000; Kluger and Wyatt, 2003). One tacit motivation seems to be the normative view that people would be better off if their beliefs conformed more closely to Bayes' Rule.

face-to-face interviews based on a scripted protocol designed to last three to 10 minutes, although no time limit was imposed. The script was visible to respondents, and the interviewer encouraged respondents to read any sample items if asked for clarification. Most interviews were collected a few yards from the registration desk at the AEA meetings that served as a passageway to the conference sessions.

The interviewer approached men who appeared at least 40 years old. An introductory statement offered respondents a choice of $3 cash or a Swiss chocolate bar, together with assurances that the survey would be short. Table 1 provides summary statistics of the survey responses used in subsequent statistical models.

Of 133 respondents, 123 (92%) said they were economists. The 10 non-economists described themselves as political scientists or academics working in fields that overlap with economics. Three quarters of respondents described their work as applied rather than theoretical. Three quarters of respondents also described their methodological approach as neoclassical (with pairwise correlation between Applied and Neoclassical of only 0.01). No respondent nonresponded when asked their age. The age distribution was remarkably symmetric, covering a large range (26 to 79) with a mean of 51 and a strong majority (119 respondents) aged 40 and above, indicating that our interviewer largely succeeded at hitting the over-40 age target.

Table 1 shows that roughly half the respondents (46%) reported having had a PSA test. Among those 50 and older, the rate of PSA testing was 65%. When asked whether they recommend that asymptomatic men in their 50s should take the PSA test as a screening for prostate cancer, most respondents (91% of the 124 who responded) responded affirmatively, with almost no difference in rates of recommendation by age. Summarized in the caption of Table 1 is information about respondents' primary subfields of specialization.

2.2 Nonresponse

The column labeled "Number of Valid Responses" shows that item nonresponse was a problem for several survey items, although not the ones we would have expected. Nine men refused to classify their work as either "more applied" or "more theoretical." And nine refused to make a recommendation about whether men in their 50s should have a PSA test. No one, however, refused to say whether he had taken a PSA.

2.3 Information Acquisition, Perceived Harms, and Information Processing

From Table 1, 22% of respondents reported having consulted written information. Only 5% reported having read a published article about PSA testing in a medical journal. The survey item labeled "Harms?" codes responses to the

	Fraction Yes	Number of valid Responses
Individual characteristics		
Keep \$3 cash?	0.12	133
Give \$3 to charity?	0.71	133
Inaccuracy measures based on absolute log deviations (and signed log deviations) of elicited beliefs	0.17	133
Economist?	0.92	133
Work is applied as opposed to theoretical?	0.75	124
Neoclassical methodological orientation?*	0.75	128
50 years old or older**	0.62	133
PSA decision and recommendation		
Did you have a PSA?	0.46	133
Would you recommend a PSA to men in their 50s?	0.91	124
Information acquisition, perceived harms, and mode of information processing		
Written info?	0.22	131
Medical journal?	0.05	131
Harms?	0.25	122
Weighed pros and cons?	0.36	128
Social influences		
Doctor influenced?	0.58	133
Spouse or relative influenced?	0.07	133
Nobody influenced?	0.15	133

	Mean Elicited Value	Std. Dev. of Mean	Number of Responses	Published Point-estimates***
Elicited frequencies				
Lifetime incidence Pr(C Lifetime)	0.27	0.019	132	0.177
Lifetime mortality Pr(D Lifetime)	0.06	0.006	132	0.028
Posterior probability Pr(C|+)	0.47	0.019	128	0.34
Sensitivity Pr(+|C)	0.72	0.018	126	0.68
Incontinence probability Pr(Incontinence|Surgery)	0.30	0.020	128	.020 to 0.29

Table 1: Survey responses

Note: *Primary subfield specializations were collected, too: 7 percent econometrics, 12 percent finance, 5 percent health economics, 7 percent economic history, 5 percent industrial organization, and 9 percent macroeconomics. No subfield indicator correlates with neoclassical methodological orientation by more than 0.12, and some, like econometrics and economic history, have slightly negative correlations with the neoclassical indicator.
**All 133 respondents reported their age in years. Mean self-reported age was 51 years old, with a strong majority (119) reporting ages of 40 or older.
***Stanford *et al.*'s (1999) NCI SEER study and Harris and Lohr (2002).

forced-choice (yes/no) question: "In your opinion are there potential harms associated with PSA screening?" The fact that only a quarter of respondents said that there were harms associated with PSA testing stands in contrast to the extensive medical literature documenting such harms (discussed below in Table 2). Only about one third of respondents reported having weighed the pros and cons about having a PSA test.

Journal	Author(s)	Comment
Archive of Internal Medicine	Concato *et al.* (2006)	"Measurement of prostate-specific antigen (PSA) in serum and digital rectal examination (DRE) are commonly used to screen for prostate cancer, yet official recommendations regarding these tests vary. For example, American Cancer Society and American Urological Association recommendations include screening for prostate cancer in men older than 50 years, using PSA testing and DRE, followed by transrectal ultrasound if either test result is abnormal. In contrast, the American College of Physicians suggests counseling regarding possible benefits and risks, and the U.S. Preventive Services Task Force (2002) found insufficient evidence to recommend screening. These positions were promulgated in the setting of data showing that the screening tests increase detection of prostate cancer but without direct evidence showing that PSA or DRE reduce mortality."
Annals of Internal Medicine	Barry (2006)	"We already know that PSA screening has a substantial downside.... The poor specificity of PSA testing results in a high probability of false positives requiring prostate biopsies and lingering uncertainty about prostate cancer risk, even with initially negative biopsy findings. Although we now know that aggressive surgical treatment of prostate cancers largely detected the "old fashioned way" without screening has a modest benefit, with about 18 cancers needing to be removed to prevent 1 death over 10 years, that benefit comes at a considerable price in terms of sexual dysfunction and incontinence. The key question is whether early detection and subsequent aggressive treatment of prostate cancers found through PSA screening prevents enough morbidity and mortality to overcome these disadvantages..."
Journal of the National Cancer Institute	Draisma *et al.* (2003)	"Whether asymptomatic men benefit from screening for prostate cancer is an unresolved question."

Table 2: Medical literature arguing against automatic PSA screening of asymptomatic men

Journal	Author(s)	Comment
New England Journal of Medicine	Steineck *et al.* (2002)	Regarding watchful waiting versus other treatment options following a diagnosis of prostate cancer, the "alternatives are associated with complex and incommensurable outcomes, and each man must judge for himself which treatment is preferable."
European Journal of Cancer	Ciatto *et al.* (2000)	"The benefits of prostate cancer screening are just theoretical, thus far unknown, and the potential risk of adverse effects much more worrying than for breast cancer: screening as a current practice is unethical, and the practice of screening, at the moment, must be limited to experimental studies." [also see Ciatto (2003) in the *British Medical Journal*]
American College of Physicians	Concato *et al.* (2006)	"Routine PSA measurement without a frank discussion of the issues involved is inappropriate."
Epidemiology	Gann (1997)	"The most important question is whether the decline in [disease-specific] mortality* will be worth the cost–in terms of anxiety, excess biopsies, and even unnecessary surgery." [also see Gann *et al.* (1995) in *JAMA*]
Journal of the American Medical Association (JAMA)	Litwin *et al.* (1995)	Regarding patients' treatment decisions and doctors' recommendations: "Little is known about how or why they make treatment decisions, how their quality of life is affected by therapy, or why physicians recommend one treatment vs. another." Regarding costs and benefits: "The traditional Western medical perspective of maximizing survival at all cost is inadequate. Indeed, the most rational approach to treating men with localized prostate cancer needs to include not only adding years to life, but also adding life to years."

Table 2: Medical literature arguing against automatic PSA screening of asymptomatic men. *Continued.*

Note: *A common recommendation of studies raising questions about PSA screening of asymptomatic men is that doctors should provide patients with information regarding pros and cons while encouraging patients to decide about PSA testing on an individual basis. Medical communication experts refer to this as the balance-sheet approach, with the goal of asking patients to weigh costs and benefits rather than making automatic decisions in favor of screening or treatment (McFall and Hamm 2003; Concato *et al.*, 2006). The National Cancer Institute (part of the U.S. National Institutes of Health) explicitly recommends against routine screening of asymptomatic men, and its website (www.cancer.gov) states that men should consider costs and benefits before deciding on a PSA test. In contrast, many hospitals and doctors have a policy of automatic screening based on age following recommendations supportive of automatic PSA screening by the American Cancer Society and the American Urological Association.

Not weighing pros and cons could, of course, be rationalized if the perceived costs or perceived benefits were zero, in which case there would be no tradeoffs to consider. When designing the survey, we worried that asking respondents if they had weighed pros and cons might not generate any variation at all, expecting nearly all economists to answer "Yes." More surprisingly, among the 30 respondents who said there were harms from PSA testing, 16 reported not weighing pros and cons. And among the 92 who said there were no harms, 30 reported having weighed pros and cons.

2.4 Elicited Frequencies

Two unconditional beliefs and three conditional beliefs were elicited:

- *lifetime incidence*: the probability that a randomly drawn male in the U.S. is diagnosed with prostate cancer within his lifetime, denoted P(C Lifetime);
- *lifetime mortality*: the probability that a randomly drawn male in the U.S. dies of prostate cancer within his lifetime, denoted P(D Lifetime);
- *incontinence probability*: the probability of incontinence conditional on surgical treatment for prostate cancer, denoted P(Incontinence | Surgery);
- *posterior probability*: the probability that an asymptomatic U.S. male in his 50s has prostate cancer conditional on a positive PSA test, denoted P(C|+);
- *sensitivity*: the probability that an asymptomatic U.S. male in his 50s has a positive PSA test conditional on having undiagnosed prostate cancer at the time of screening, denoted P(+|C).

The last five rows in Table 1 report mean subjective beliefs and corresponding point estimates published in the *National Cancer Institute's Surveillance Epidemiology and End Results [SEER]* database (Stanford *et al.*, 1999) and *Annals of Internal Medicine* (Harris and Lohr, 2002). Respondents' beliefs about these five probabilities tended to be slightly too large but not terribly inaccurate with respect to the published point estimates.

2.5 Recent Shifts in Expert Opinion That Make PSA Testing an Important Decision to Study

Before introducing measures of consistency and accuracy of beliefs, Table 2 summarizes eight frequently cited medical studies about the risks and benefits of PSA testing with quotations that highlight recent shifts in expert opinion. In contrast to policies in place among many clinicians and hospitals, the US Preventative Services Task force currently recommends against PSA screening (see http://www.uspreventiveservicestaskforce.org/Page/Topic/recommendation-summary/prostate-cancer-screening). Instead of automatic screening for all men once they reach 40, the recommendation of the National Institutes of Health (NIH) is that men weigh the pros and cons of PSA testing and make their decision on an individual basis in consultation with their doctor. Table 2 motivates the use of PSA testing as a potentially high-stakes decision of interest to decision theory because of the divergence between expert recommendations (against routine screening of asymptomatic men) and the common clinical practice of recommending testing for all men once they reach a particular age.

After gaining FDA approval in 1986 for use among men already diagnosed with prostate cancer, PSA testing spread rapidly as a screening tool for asymptomatic men. By the late 1990s, as many as half of American men over the age of 50 were estimated to have undergone PSA testing (Gann, 1997). Aside from the large direct costs of financing mass screening, estimated at $12 to 18 billion per year (U.S. Preventive Services Task Force, 2002, p. 128), another point of contention concerns the benefits of early detection (Stanford *et al.*, 1999; U.S. Preventive Services Task Force, 2002). Most prostate cancers are slow growing. A large majority of men with prostate cancer die of other causes first. And benefits of early detection may be limited in the case of fast-growing cancers, too, insofar as treatments have poor rates of success. Although some studies report that early detection of prostate cancer reduces disease-specific mortality, there is no evidence demonstrating that early detection reduces overall mortality (Ciatto *et al.*, 2000; Holmberg *et al.*, 2002; Yao and Lu-Yao, 2002; Draisma *et al.*, 2003; Concato *et al.*, 2006). Recent randomized trials in the U.S. also find no evidence that PSA screening reduces death from prostate cancer or death from cancer in general; mortality rates were slightly higher in the group that underwent screening (Andriole *et al.*, 2009). Compared to this ambiguous evidence about the benefits of PSA testing, the evidence of harms is relatively clear. Harms from prostate cancer screening include psychological stress, needless biopsies following false positives, and overtreatment of nonlethal prostate cancers, resulting in complications such as incontinence and impotence (Wang and Arnold, 2002; Hawkes, 2006).

2.6 Elicitation Technique for Measuring Accuracy and Consistency

We sought to construct measures of accuracy and logical inconsistency that rely on non-overlapping sets of survey items so that these two measures of the belief quality do not functionally depend on each another. Our interview script first elicits two unconditional beliefs:

> The main focus of the survey is prostate cancer and PSA (Prostate Specific Antigen) screening. I won't ask any personal questions

> about the illness itself, just about screening. I'd like to elicit your best guesses about the risks of prostate cancer. For a randomly drawn American male, I'd like you to guess the probability that he will *be diagnosed* with prostate cancer in his lifetime? What would you say the probability is that he will die from prostate cancer in his lifetime?

The unconditional beliefs elicited by the proceeding questions are referred to as lifetime incidence and lifetime mortality, denoted $\mathrm{P(C\ Lifetime)}_i$ and $\mathrm{P(D\ Lifetime)}_i$, respectively. The difference between these unconditional beliefs and published point estimates provide the basis for individual-level measures of inaccuracy as defined in the next subsection.

The interview script proceeds by eliciting two conditional beliefs that we use to measure inconsistency of beliefs:

> Now I'm going to ask you about American males in their 50s who have no symptoms, have never been diagnosed with prostate cancer, and are screened with a PSA test for the very first time. One leading study suggests that 5% of randomly sampled men from this population have a positive PSA. It's also estimated that 2.5% actually have prostate cancer at the time of screening, which includes those whose PSAs failed to detect the disease.[6] [source: Harris and Lohr, 2002, *Ann Intern Med*]. Given a positive PSA, I'd like you to estimate the probability that a man actually has prostate cancer. And given cancer at the time of screening, what would you say the probability of a positive PSA is?

[6] One may question whether the phrase, "which includes those whose PSAs failed to detct the disease," is leading language that could bias conditional beliefs elicited using this interview script. The reasoning behind including this phrase was our view that nearly everyone knows that screening tests have imperfect sensitivity (i.e., $P(+|C) < 1$ is common knowledge), and when providing respondents with the published unconditional probability $P(+) = 0.05$, we wanted to make sure that they knew it was an *unconditional* probability (including both men with and without prostate cancer). The literature on risk communication and doctors' understanding of the statistical properties of both PSA testing and mammography screenings reveals persistent problems with false positives in particular, which suggests there is a very real asymmetry in people's understanding of type-1 and type-2 errors in the context of disease screening. There is considerable evidence that doctors and patients alike routinely under-appreciate false positives (Gigerenzer *et al.*, 2007). To address this concern over asymmetric language, we can check the distance between mean conditional beliefs and their objective values as well as compare their standard deviations to see if there is evidence that the survey instrument led respondents to be relatively more advantaged at calculating $P(C|+)$ than at calculating $P(+|C)$. The mean subjective belief for $P(C|+)$ of 0.72 (or 72%) is closer to its objective value of 0.69 (from Table 1) than the mean subjective belief for $P(C|+)$ of 0.47 is to its objective value of 0.34. Similarly, there one observes no gross difference in the conditional belief variables' standard deviations ($std(P(+|C)_i = 20.0$ and $std(P(C|+)_i = 21.8$), which are very close and ordered opposite from what would be predicted under the hypothesis that language in the elicitation led or primed our respondents.

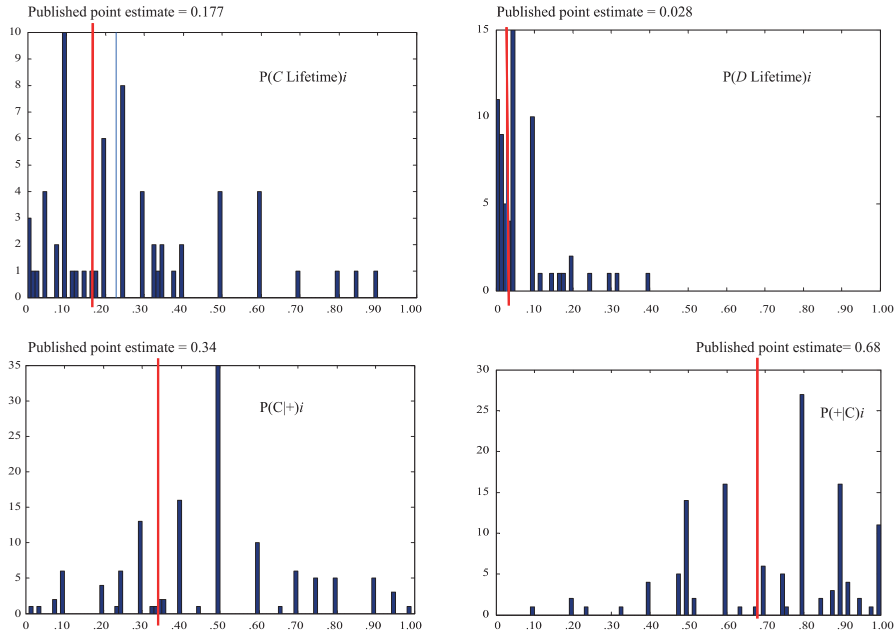

Figure 1: Elicited belief distributions.

The resulting conditional beliefs — the probability of prostate cancer conditional on a positive PSA test, denoted $\mathrm{P(C|+)}_i$, and the probability of a positive PSA test conditional on cancer, denoted $\mathrm{P(+|C)}_i$ — provide the basis for measuring non-Bayesian inconsistency as defined below.

Applying the definition of conditional probability and substituting in the two unconditional probabilities that were provided to respondents results in a restriction on the ratio of elicited conditional beliefs: $\mathrm{P(C|+)}_i/\mathrm{P}(+|C)_i = P(C)/P(+) = 2.5/5 = 1/2$. Respondents may know nothing about the relevant medical studies and published PSA facts yet conform perfectly to Bayes' Rule. In fact, there are infinitely many pairs of conditional beliefs that conform perfectly to the ratio restriction above regardless of whether $\mathrm{P(C|+)}_i$ and $\mathrm{P(+|C)}_i$ are near or far from published estimates of those conditional probabilities (which were not provided to respondents). Figure 1 shows the elicited belief distributions.

2.7 Inconsistency and Inaccuracy

We consider measures of inconsistency based on deviations of the ratio of elicited conditional beliefs, $\mathrm{P(C|+)}_i/\mathrm{P}(+|C)_i$, from the ratio restriction that Bayes' Rule imposes, $\mathrm{P(C)/P(+)} = 1/2$. Similarly, we consider measures of inaccuracy based on the deviations of elicited unconditional beliefs from their corresponding published point estimates in Table 1. Differences in levels, percentage deviations, and log-approximated percentage deviations using both signed and absolute versions of those deviations were analyzed.

Absolute log-approximated percentage deviations from the Bayesian ratio restriction provide the following measure of inconsistency:[7]

$$inconsistency_i = |\log[\mathrm{P(C|+)}_i/\mathrm{P(+|C)}_i] - \log[1/2]|.$$

Absolute log-approximated percentage deviations with respect to published point estimates in Table 1 provide the following measure of inaccuracy:

$$inaccuracy_i = (|\log[\mathrm{P(CLifetime)}_i/0.177)]| + |\log[\mathrm{P(D\ Lifetime)}_i/0.028]|)/2.$$

This definition averages deviations of beliefs about lifetime incidence and lifetime mortality.[8]

Figure 2 presents a scatter plot of $inaccuracy_i$ and $inconsistency_i$. The 24 individuals clustered along the y-axis (with $inconsistency_i = 0$) are Perfect Bayesians in the sense that their conditional beliefs conform perfectly with the ratio restriction imposed by Bayes' Rule. We note that the two most inaccurate individuals in the sample (northern-most observations along the y-axis plotted in Figure 2) turn out to be Perfect Bayesians. In contrast, the two most inconsistent individuals (eastern-most) have below-average inaccuracy and are well inside the lower half of Figure 2 containing observations (with inaccuracy below the midpoint of its range of variation). The bivariate data (without conditioning on other observable features) do not suggest there is empirical convergence of these two normative criteria. Further analysis of inaccuracy and inconsistency in the presence of conditioning information in other survey items summarized in Table 1 also fails to uncover any positive association between inconsistency and inaccuracy.

---

[7] All data analysis reported in this paper was repeated using alternative definitions of inconsistency and inaccuracy based on other functional specifications of the deviation.For example, deviations can be measured in *percentage points* (although it gives disproportionate influence to respondents with large-magnitude beliefs): $|\mathrm{P(+|C)}_i - 2\mathrm{P(C|+)}_i|$. Another deviation we considered was raw percentage deviations rather than log approximations: $|[\mathrm{P(C|+)}_i/\mathrm{P(+|C)}_i - 1/2]/(1/2)|$, which produces a more spread-out distribution and is not invariant to algebraically equivalent re-statements of the restriction such as $|[\mathrm{P(+|C)}_i/\mathrm{P(C|+)}_i - 2]/2|$. Dichotomization only strengthens the case for our interpretations (see Table 3 below).

[8] Lifetime incidence and lifetime mortality are used because the conditional beliefs were already used to compute inconsistency. Most of the variation in *inaccuracy* derives from beliefs about mortality, which is rarer than incidence and therefore generates a wider range of percentage deviations. We re-ran all data analysis using alternative measures of *inaccuracy*: lifetime incidence deviations alone, lifetime mortality deviations alone, and an average of five deviations based on all five beliefs, revealing no positive correlations with *inconsistency*.

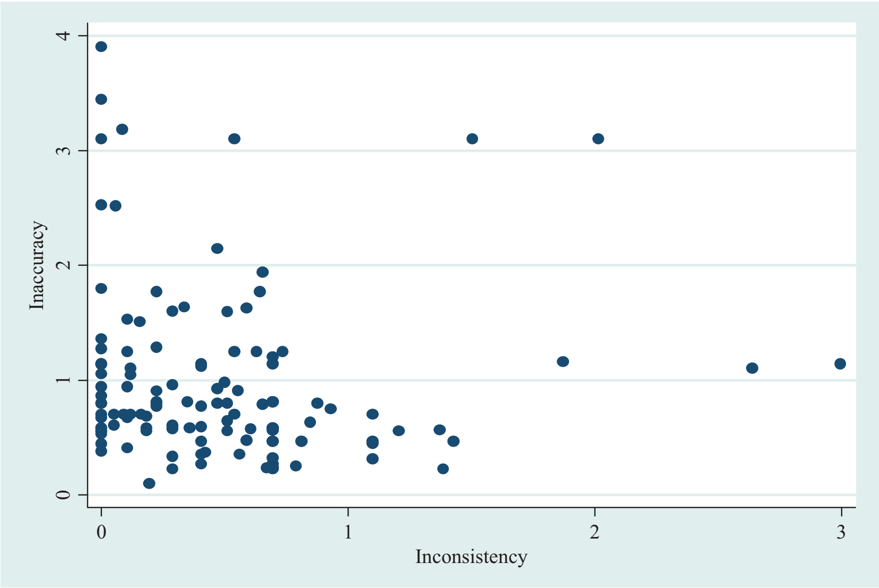

Figure 2: Inconsistency and inaccuracy ($N = 125$).

Note: The bivariate regression line is E[inaccuracy|inconsistency] $= 1.00 - 0.06 *$ inconsistency. Because inconsistency and inaccuracy are defined as log deviations, the coefficient -0.06 ($p = 0.645$) can be interpreted as the elasticity of inaccuracy (percentage-point deviations from published incidence and mortality rates) with respect to inconsistency (percentage-point deviation from Bayes Rule). Pairwise correlation is -0.042.

2.8 Accuracy Contrasts between Perfect Bayesians and non-Bayesians

Table 3 presents four binary contrasts of mean inaccuracy among dichotomized subsamples according to belief consistency. The units are log-approximated percentage deviations from published point estimates on a decimal scale (e.g., a difference of 0.1 is approximately 10 percentage points). The four contrasts in Table 3 and corresponding t statistics are, of course, not independent because they use overlapping observations, dichotomized using different thresholds to compare more Bayesian versus less Bayesian subsamples. These subsamples are defined as: Perfect Bayesians ($\textit{inconsistency} = 0$) versus non-Bayesians ($\textit{inconsistency} > 0$); below-median versus above-median *inconsistency*; bottom versus upper quartiles of *inconsistency*; and Near Bayesians (an inclusive classification for anyone whose inconsistencies can be modeled as Bayesian beliefs plus a noise term) versus so-called Emersonians (explained below) who commit gross errors in conditional probabilistic reasoning.[9]

[9]The label refers to Emerson's (1841) "Self Reliance," in which Emerson wrote: "The other terror that scares us from self-trust is our consistency ... A foolish consistency is the

		Consistent	Inconsistent		Consistent	Inconsistent		Consistent	Inconsistent		Consistent	Inconsistent	
	Pooled Mean	24 Perfect Bayesians	101 Deviators from Bayes	t stat	60 Strictly Below Median inconsistency	65 Weakly Above Median inconsistency	t stat	36 Weakly Below 25th Percentile inconsistency	34 Weakly Above 75th Percentile inconsistency	t stat	80 Near Bayesians	45 Emersonians	t stat
Inaccuracy measures based on absolute log deviations (and signed log deviations) of elicited beliefs													
inaccuaracy	0.99	1.26	0.90	1.7	1.08	0.87	1.6	1.26	0.77	2.5	1.08	0.78	2.2
signed inaccuracy	0.01	−0.56	0.16	−2.2	−0.12	0.15	−1.3	−0.45	0.04	−1.6	−0.14	0.32	−2.1
Log deviations													
log (incidence/ 0.177)	−0.06	−0.43	0.08	−1.7	−0.13	0.09	−1.0	−0.44	−0.04	−1.3	−0.11	0.15	−1.2
log (mortality/ 0.028)	0.07	−0.69	0.23	−2.2	−0.11	0.21	−1.2	−0.48	0.11	−1.5	−0.18	0.48	−2.5
log (posterior/ 0.34)	0.18	0.00	0.22	−2.1	0.11	0.23	−1.1	0.09	0.12	−0.2	−0.11	0.67	−7.9
log (sensitivity/ 0.64)	0.06	0.06	0.07	0.0	0.11	0.02	1.5	0.13	−0.10	2.5	0.09	0.01	1.3
Measures of inconsistency													
inconsistency	0.48	0.00	0.59	–	0.12	0.81	–	0.03	1.05	–	0.34	0.73	–
signed inconsistency	−0.17	0.00	−0.21	–	−0.06	−0.28	–	−0.02	−0.28	–	0.14	−0.73	–

Table 3: Contrasts in mean inaccuracy among consistent and inconsistent subsamples

Note: *Inaccuracy is the (within-individual) simple average of the four absolute log deviations. Signed inaccuracy is the simple average of those same log deviations without taking absolute values.
**Inconsistency is the absolute percentage error of the elicited ratio, *sensitivity/posterior*, relative to the correct ratio of 2. Signed inconsistency is the same as inconsistency but without absolute values.

The rows of Table 3 contain mean values of inaccuracy, signed inaccuracy (removing the absolute value operation in the definition of inaccuracy presented earlier), four log deviations corresponding to each belief (with no averaging), and mean values of inconsistency and signed inconsistency within each subsample. Signed inaccuracy and inconsistency allow deviations with opposite signs to cancel (to some extent) when summing over individuals.

Reading horizontally across the first row, Table 3 shows that the average Perfect Bayesian (among 24 individuals with $inconsistency_i = 0$) is more inaccurate than the rest of the sample (1.26 versus 0.90). Similarly, the second contrast of subsamples with below- versus above-median inconsistency shows that the lower half of the inconsistency distribution has greater inaccuracy than the upper half (1.08 versus 0.87). In the third contrast, the lower quartile of the inconsistency distribution has greater inaccuracy than the upper quartile (1.26 versus 0.77). And in the fourth contrast of Near Bayesians versus Emersonians, accuracy is, once again, negatively associated with consistency: mean inaccuracy of 1.08 among Near Bayesians versus 0.78 among Emersonians.

The second row of Table 3 shows that the beliefs of consistent respondents tend to be too small, whereas the beliefs of inconsistent individuals tend to be too large. Consistent individuals' beliefs are not, however, generally closer to the published point estimates. Rows 3 and 4 show mean log deviations for lifetime incidence and mortality. These disaggregated bivariate contrasts reveal no general tendency for consistent individuals to have more accurate beliefs regardless of which threshold is used to dichotomize the sample.

2.9 Taxonomy of Inconsistencies: Emersonians and Near Bayesians

Closer examination of the elicitation scheme reveals distinct ways in which a respondent can deviate from Bayes' Rule. Some respondents are within plausible bounds (defined below) and could be modeled as if their beliefs were Bayesian with an error term (referred to as Near Bayesians). Other respondents' beliefs commit more fundamental violations of probabilistic logic that are more difficult to interpret as noisy Bayesian beliefs (referred to as Emersonians).

We define three types of gross violations of probabilistic reasoning, any one of which would indicate a belief generating process that cannot be easily reconciled with the definition of conditional probability. The first gross logical error is $\mathrm{P(C|+)}_i > 0.50$. The definition of conditional probability states that $\mathrm{P(C|+)} = \mathrm{P(C \cap +)}/P(+)$. The numerator refers to an intersection of events with an obvious upper bound: $\mathrm{P(C \cap +)} \leq \min\{\mathrm{P(C), P(+)}\} = 0.025$. The unconditional probabilities provided to respondents therefore imply that ratio of conditional beliefs must be bounded above by 1/2: the upper bound of

hobgoblin of little minds, adored by little statesmen and philosophers and divines. With consistency, a great soul has simply nothing to do."

$P(C \cap +) \leq 0.025$ divided by the value of $P(+) = 0.05$ provided in the interview script implies that $P(C|+)_i \leq 0.025/0.05 = 1/2$. Beliefs at the upper bound of 1/2 correspond to the belief that there are no false positives. Of 133 respondents, 36 (34 economists and 2 non-economists) violated this logical bound.

The second gross departure from probabilistic logic is $P(C|+)_i > P(+|C)_i$. The definition of conditional probability implies that the numerators of $P(C|+) = P(C \cap +)/P(+)$ and $P(+|C) = P(C \cap +)/P(C)$ are identical, while the denominators are known unconditional probabilities. The given information, $P(C) = 0.025$ and $P(+PSA) = 0.05$, should imply that $P(C|+)_i \leq P(+|C)_i$ for all beliefs about the intersection, $P(C \cap +)_i$, holding with equality only when $P(C \cap +)_i = 0$. Eleven respondents strictly violated this condition, 9 of whom also committed the first gross departure from probabilistic reasoning.

The third logical error is $P(C|+)_i = P(+|C)_i$. As long as there is at least one man whose prostate cancer is correctly identified by a PSA test (i.e., $P(C \cap +) > 0$), then $P(C|+PSA)$ cannot be zero, which implies that the inequality $P(C|+) < P(+|C)$ must hold strictly. Sixteen respondents provided equal conditional beliefs. Of these, seven also violated the first logical restriction, and seven others violated the second restriction. In total, 45 respondents committed at least one of the three errors resulting in the designation *Emersonian.*

2.10 Perceived accuracy of the PSA test

One final comparison is considered regarding the perceived versus objective overall accuracy of the PSA test. Accurate PSA tests occur when the test is positive and a man has prostate cancer (with associated probability $P(+ \cap C)$) or when the test is negative and a man does not have prostate cancer (with associated probability $P(- \cap \sim C)$). The following calculation expresses the probability that the PSA test is accurate as a function of the conditional probabilities corresponding to the conditional belief data ($P(C|+)_i$ and $P(+|C)_i$), which affect the probability of (the complementary event of) an inaccurate PSA test, $P(- \cap C) + P(+ \cap \sim C)$ corresponding to false negatives and false positives:

$$\begin{aligned} P(\text{test is accurate}) &= 1 - P(- \cap C) - P(+ \cap \sim C) \\ &= 1 - P(-|C)P(C) - P(\sim C|+)P(+) \\ &= 1 - (1 - P(+|C))P(C) - (1 - P(C|+))P(+). \end{aligned}$$

The variable *perceived accuracy* is estimated by substituting each individual's conditional beliefs $P(C|+)_i$ and $P(+PSA|C)_i$ for $P(+|C)$ and $P(C|+)$ in the equation directly above and using the published values $P(C) = 0.025$ and $P(+) = 0.050$. The mean value of *perceived test accuracy* is 0.973 with standard deviation = 0.013 and an empirical range of 0.937 to 0.999. The objective probability that the test is accurate, computed using the published point estimates for P(+|C) and P(C|+) in Table 1, is: $1 - (1 - 0.68) \times 0.025 - (1 - 0.34) \times 0.050 = 0.959$. The bivariate regression coefficient on *inconsistency* regressed on *perceived test accuracy* is -0.13 ($p = 0.575$). Similar to earlier findings, the beliefs of the two most inconsistent individuals are again very well-calibrated to the objective accuracy of the PSA test.

3 Conditional Effects of Consistency on Belief Accuracy and PSA Test Taking

If deviations from Bayes' Rule were a good predictor of economic loss, then we would expect to see inconsistency with respect to Bayes' Rule affect either the objective accuracy of men's beliefs or the actions that they choose to take (i.e., the conditional probability of having a PSA test). Further analysis using a loss function framework faces at least two challenges, however. The first challenge is to describe in sufficient detail the states of nature over which losses would need to be integrated when computing expected loss (i.e., risk). The states of nature would consist of a large number of pathways that combine possible screening decisions, diagnoses, treatments, and outcomes along both the C and ∼C branches in the extensive-form event tree. A second challenge would be to account for men's different valuations, perceived effectiveness of treatments, and perceived likelihoods of outcomes.

As a partial step, this section reports regression results (extending the bivariate results reported in the previous section), which provide tests for the effects of inconsistency on inaccuracy and decisions about PSA testing in the presenence of controls. One important limitation is that our survey data do not provide detailed controls measuring men's beliefs about the effectiveness of treatments along different branches of the event tree mentioned above. We do, however, use beliefs about the probability of incontinence in the event that prostate cancer is treated with surgery as a partial control. The goal of conditional testing is to detect evidence that *inconsistency* is associated with either *inaccuracy* or actions (based on beliefs about prostate cancer and the PSA test) that could be interpreted as connecting non-Bayesian beliefs to economic loss through one of these two channels (belief accuracy or actions).

3.1 Does Inconsistency Affect the Expected Inaccuracy of Beliefs?

We discuss (without reporting the full set of regression results) the effect of inaccuracy on inconsistency in the presence of a full set of controls: having consulted written information, information processing (i.e., weighing pros and cons), social influencers, a quadratic function of age, other individual characteristics from the survey, and subfield indicators. The effect of *inconsistency* on *inaccuracy* turns out to be little changed from the bivariate regression line in Figure 2. The regression coefficient on inconsistency was -0.06, $p = 0.645$ in the bivariate model, and -0.08 ($p = 0.550$) in the conditional model. Similarly, for every intermediate specification involving different subsets of the regressors, we never observed a positive coefficient suggesting a positive association between consistency and accuracy.

3.2 Does Inconsistency Affect the Probability of PSA Testing?

Table 4 presents estimates of four linear probability models of binary PSA test decisions and t statistics computed using robust standard errors.[10] The *fundamental model* assumes that PSA decisions are a function of the five subjective beliefs (proxying for beliefs about risks of prostate cancer and net benefits of PSA testing) and a quadratic function of age. The *add info processing model* includes individual variation in information acquisition, information processing, and inconsistency, in addition to all variables in the *fundamental model*. Finally, the *add influencers model* (again encompassing previous models) allows the probability of PSA testing to depend on social influencers. The first three models use the binary PSA decision as the dependent variable and the fourth model uses binary PSA recommendations as the dependent variable to investigate whether the conditional information in the encompassing model can explain the large gap between unconditional mean rates of PSA decisions and recommendations, 46 versus 91%, respectively.

We find statistical confirmation of economists' self-reports that most do not weigh costs and benefits when deciding whether to have a PSA test. Across all three models, the individual belief variables in the first five rows of Table 4 have surprisingly weak effects on the probability of having PSA testing. For example, the perceived risk of incontinence, which one might have guessed would strongly condition the likelihood of PSA testing, has (at most) very modest effects: the coefficients on log(incontinence/0.150) imply that a man who perceives the risk of incontinence as being twice as large as the average man does is, at most, 6 to 8 percentage points less likely to have a PSA. Coefficients on information acquisition and processing (i.e., pros-and-cons deliberation and logical inconsistency) are nowhere large or statistically significant.

In the *fundamentals model*, the joint test that the five belief variables all have zero coefficients corresponds to the hypothesis that subjective beliefs about cancer risks and the benefits of treatment do not influence PSA decisions.

[10]Logit and probit models produce qualitatively identical results. Similar to Wisdom *et al.* (2010), we use the linear probability model estimated by OLS (with robust standard errors) to provide easy-to-interpret magnitudes of estimated effects on binary outcomes (healthy versus unhealthy menu choices, in their case, and PSA decisions in ours). The linear probability model has the advantage of easily correcting for heteroscedasticity of errors. We checked that none of the important effect sizes or qualitative results change in the logit and probit specifications of the empirical model.

	Three PSA Decision Models:							
	fundamental		*add info-processing*		*add influencers*		PSA Recommendation	
predictors	coefficient	t	coefficient	t	coefficient	t	coefficient	t
log(incidence/0.177)	0.05	1.0	0.07	1.4	0.04	0.9	−0.11	−2.4
log(mortality/0.028)	−0.01	−0.3	0.00	0.1	0.01	0.3	0.10	2.8
log(posterior/0.34)	−0.09	−1.6	−0.06	−0.9	−0.05	−0.7	−0.05	−0.7
log(sensitivity/0.64)	0.10	1.0	0.14	1.2	0.16	1.5	0.18	1.4
log(incontinence/0.150)	−0.06	−1.6	−0.07	−1.7	−0.08	−2.3	−0.07	−2.7
age	−0.03	−1.1	0.00	0.1	−0.02	−0.6	0.02	0.7
age squared	0.00	2.0	0.00	0.6	0.00	1.3	0.00	−0.7
cash?(1/0)			−0.15	−1.5	−0.17	−2.0	−0.10	−0.9
chocolate?(1/0)			−0.08	−0.7	−0.09	−0.8	−0.08	−0.9
procon?(1/0)			−0.06	−0.6	−0.04	−0.4	−0.05	−0.6
consult written?(1/0)			0.14	1.5	0.15	1.6	0.13	1.4
inconsistency			0.01	0.2	0.00	0.1	−0.02	−0.3
nobody influenced?(1/0)					−0.09	−0.7	−0.17	−1.3
doctor influenced?(1/0)					0.27	2.9	−0.03	−0.3
constant	0.79	1.0	−0.09	−0.1	0.52	0.6	0.65	0.8
R_2	*0.34*		*0.38*		*0.46*		*0.18*	
test stat for H0*	*0.00*						*0.00*	
Pr(test stat>observed | H0)	*0.13*		*0.14*		*0.03*		*0.01*	
Sample Size	*121*		*114*		*114*		*114*	

Table 4: Linear probability models of PSA decisions and recommendation

Note: *H0 is the joint hypothesis that the first five variables, which proxy for perceived costs and benefits, have zero effect on the probability of having (or recommending) a PSA. The test statistic is distributed as F(5, sample size minus number of regressors) under the null.

The second-to-last row in Table 4 shows p-values corresponding to tests of that joint hypothesis. Those tests demonstrate the weak joint explanatory power of subjective beliefs in the first two models. This weak explanatory power does not result from overall weakness of the prediction equation, however, because likelihood ratio tests in all models easily reject the hypothesis that all coefficients in the model are zero. The p-value for the joint null in the *add influencers* model suggests that beliefs about the costs and benefits of PSA testing do have significant explanatory power once information about social influencers is included. The *doctor influenced* variable reveals strong conditional correlation between doctor's influence and taking the PSA test despite the obvious incentive mismatch in doctor-patient interaction, leading to well-documented problems of defensive medicine, over-diagnosis, over-prescription, overtreatment, and other potential problems that economists are well aware of (see Behrens *et al.*, 2005; Loewenstein, 2005; Studdert *et al.*, 2005; and Sorum *et al.*, 2004, for more on doctor-patient incentive mismatch).

3.3 PSA Recommendation

Pairwise correlation between PSA recommendations and self-reported decisions is surprisingly small (0.09) and far from statistical significance. As mentioned above, the unconditional rate of recommendation is double the rate of PSA test taking. To keep the sample size the same as the other models in Table 4, the PSA recommendation variable was modified to a forced-choice version coding nonresponses as zeros. Even in this forced-choice version with a conservative default rule for nonresponse, the rate of recommendation remains nearly twice as large as the rate of PSA taking: 46 versus 85%. Beliefs have more predictive power for PSA recommendations than for PSA decisions. Inconsistency plays a very limited role.

4 Discussion

4.1 Summary

This study elicited subjective belief data providing measures of both consistency with respect to Bayes' Rule and accuracy with respect to objective frequencies. These measures of inconsistency and accuracy revealed no positive (and often negative) correlations. Bayesian consistency (i.e., conditional beliefs that conform to the definition of conditional probability) and objective accuracy of beliefs (i.e., corresponding, or being well calibrated, to objective frequencies) are both theoretically and empirically distinct as normative criteria.

Our elicitation technique for belief inconsistency provided participants with two unconditional probabilities based on published medical studies and then elicited conditional beliefs whose ratio is constrained to equal the ratio of known unconditional probabilities by the definition of conditional probability. Individuals (even economists who are well equipped to apply Bayes' Rule) vary considerably in the extent to which their conditional beliefs conform to restrictions imposed by the logic of probability theory.

A second goal of this paper was to organize the belief data to test for evidence that inconsistent beliefs might cause economic losses. While acknowledging the problem of controlling for men's valuations and perceived likelihoods of different treatment outcomes, we argued that at least a partial step toward testing whether inconsistency is costly would be to examine two potential mechanisms linking inconsistency to economic loss: that inconsistency leads to inaccuracy, or that inaccuracy affects PSA testing decisions (holding the beliefs about risks of prostate cancer and net benefits of PSA testing constant). The data reveal no evidence of economically or statistically significant effects through either channel.

4.2 Which Normative Criterion Fits the Environment?

When evaluating belief data, social scientists sometimes tacitly assume that belief consistency and belief accuracy ought to be in harmony even though, analytically and empirically, they may be unrelated or negatively correlated. If there are high-stakes decision environments that reward accuracy but not consistency, then it would be unsurprising to find that people with consistent beliefs are no more or less likely to have accurate beliefs. Absent any evidence linking consistency and accuracy, those making normative claims or suggesting institutional designs that aim to improve belief rationality by the sole criterion of *consistency* would do well to delimit those normative judgments and tailor any proposed nudges to match the environments in which it has been confirmed that consistency is in fact rewarded. In environments that primarily reward accuracy but not consistency, normative and prescriptive analysis could perhaps do better by focusing directly on enabling improvements in accuracy rather than worrying about the intermediate step of checking for, or encouraging, belief consistency, which may not matter in a particular classes of decision problems.

4.3 Social Heuristics in Medical Decision Making

With the usual caveats required when interpreting self-reports about issues as personal as medical decision making, we asked respondents how much written information they had acquired, the sources of that information, and whether or not they had weighed the pros and cons when deciding whether to have a PSA test. More than half said that they had not weighed pros and cons. One may wonder whether these data are simply too noisy to reveal the underlying mechanisms that would otherwise exhibit positive associations between consistency and accuracy. We argue, on the contrary, that respondents' self-reported PSA decisions become intelligible with acceptable levels of model fit under the alternative hypothesis that economists, like many people, sometimes rely on the simple heuristic of following doctors' advice, sometimes referred to as a *white-coat heuristic*, that is, when in a hospital or at a doctor's office, people adhere to what doctors (usually in white coats) recommend (Wegwarth and Gigerenzer, 2013). The social influencer indicator variables, especially *doctor influenced*, add considerable explanatory power to the conditional models in Table 4.

There is abundant evidence that incentive mismatch between doctors and patients can lead to defensive medicine (i.e., treatments provided for the doctor's benefit of legal protection) and overtreatment of cancers that would not have caused death. There is also abundant evidence documenting large gaps in doctors' statistical literacy and their knowledge of research and statistical evidence (Gigerenzer *et al.*, 2007). On the other hand, the time costs of accessing information about prevalence and mortality of prostate cancer, together with evidence-based recommendations on screening and treatment, would amount to little more than a few mouse clicks as this information is readily available online (e.g., the U.S. Preventive Services Task Force online database). An expected utility maximization model whose solution is an action rule that relies solely on doctors' advice without conditioning on other sources of information would require strong restrictions on functional forms in order for patients' subjective beliefs about risks of cancer and PSA testing to not influence a man's probability of having a PSA test.

4.4 Interpretations

Why would smart people hold inconsistent subjective beliefs? Gilboa *et al.* (2008) provide examples of decision contexts (e.g., wars, or a coin that one has never seen or flipped before) in which they argue it would be irrational to hold probabilistic beliefs. According to them (and others), non-standard reasoning processes that generate behavior inconsistent with axioms of internal consistency can be defended and, in some contexts, shown to have advantages over decision processes that adhere strictly to consistency (e.g., Gilboa and Schmeidler, 1995; Samuelson, 2001; Aragones *et al.*, 2005; Spiegel *et al.*, 2007; Robson and Samuelson, 2009; Bardsley *et al.*, 2010). Grunwald and Halpern (2004) identify a related problem in which non-Bayesian updating provides more precise predictions. In both theoretical and empirical studies, less-is-more effects by which non-standard beliefs and heuristics that ignore relevant information are shown to provide real economic benefits and improvements in predictive accuracy (e.g., Hogarth and Karelaia, 2005, 2006; Baucells *et al.*, 2008; Berg and Hoffrage, 2008; Goldstein and Gigerenzer, 2009).

Sugden (1991) argues against the normative interpretation of expected utility theory, and Starmer's (2000, 2005, 2009) historical and methodological analyses of normative debates about Bayesian reasoning and expected utility theory arrive at similar conclusions. Camerer and Hogarth (1999) suggest that learning about the consequences of one's inconsistency occurs relatively slowly, and Loewenstein (1999, 2005) argues that many high-stakes decisions, especially medical decisions, are one-shot (without repetition in decision makers' health decision-making environments). These findings raise questions about whether it is reasonable to assume that inconsistency should be competed away or reduced as the result of experience (*c.f.*, Braga *et al.*, 2009). In high-stakes decisions (e.g., medical decisions with substantial mortality risk, financial decisions involving a large fraction of one's wealth, or career and relationship advice among loved ones), many who are well-equipped to follow axiomatic requirements of consistency nevertheless choose to apply normative criteria beyond, or in conflict with, consistency.[11]

4.5 Decision-Making Process in PSA Testing

Table 1 showed that only 46 out of 128 respondents reported having weighed pros and cons when deciding on PSA testing, including 16 who did not weigh pros and cons despite having reported that they perceived potential harms. This suggests a thought process in line with Gilboa *et al.*'s (2009) "view of rationality that requires a compromise between internal coherence and justification." Social influencers provide justification in the social environments in which people commonly make medical decisions (e.g., having the PSA test because a spouse or doctor told me to do so, or because someone I know said to).

4.6 Guess-50 Heuristic

Respondents may have simply guessed "50%" when facing elicitation of beliefs about probabilities for which they had only agnostic priors. We coded the number of times respondents guessed 50 to see if uninformed priors indicated by guessing 50% was correlated with either consistency or accuracy. Among the five elicited beliefs, the maximum number of times anyone in the sample guessed 50 is twice. Interestingly, the 22 individuals who guessed 50 twice had more accurate beliefs (mean inaccuracy of 0.71, SE = 0.01) than those

[11] According to reliable sources, a well-known decision theorist and proponent of strictly normative interpretations of axiomatic decision theory faced the decision of whether to take a job offer from a competing university. He deliberately chose to ignore normative decision theory based on consistency axioms. When colleagues asked him why he did not simply choose a prior, compute the expected utilities associated with each job offer, and then choose the action with maximal expected payoff, the decision theorist responded in exasperation: "Come on, this is serious!" (Gigerenzer, 2004).

who never guessed 50 (mean accuracy of 1.02, SE = 0.09). Two of 24 Perfect Bayesians guessed 50 twice (e.g., $P(C|+)_i/P(+|C)_i = 50/100$ or 25/50 would allow for guessing 50 and being perfectly Bayesian). Emersonians and Near Bayesians guessed 50 at roughly the same rates. And inconsistency was uncorrelated with guessing 50.

4.7 Additional Evidence Regarding Social Influences on PSA Decisions

There is a large difference in rates of PSA taking between those who reported that nobody influenced them and those who reported at least one influencer: 36 versus 78%. No other variable in our data has such a large bivariate association with PSA taking. This (singularly) strong bivariate leverage suggests that social influence may completely override cost-benefit thinking, complementing regression evidence in Table 4 pointing to the importance of social influencers. For example, there is only a modest 15 percentage-point difference in rates of PSA taking between respondents who weighed pros and cons (76%) and those who did not (61%), and this difference disappears within the subsample of those reporting having been influenced by at least one other person (most commonly, a spouse).

4.8 Why Economists?

To improve the chances of finding empirical links between logical consistency and the objective accuracy of beliefs, the data were collected mostly from economists. Gaechter *et al.* (2009) argue that empirical findings of anomalous behavior by economists are especially convincing because one would expect economists' professional training to limit algebraic and statistical errors while providing unusually strong awareness of psychological mechanisms thought to give rise to anomalies. Our sample size of 133 was comparable to theirs, which was 120. Previous studies have shown that economists behave differently from non-economists because of both selection and training (Carter and Irons, 1991; Frank *et al.*, 1993; Yezer *et al.*, 1996). Surveys of economists have shown that economists' statistical reasoning and policy views differ substantially from those of non-economists (Caplan, 2001, 2002; Blendon *et al.*, 1997). Also relevant to the medical decision-making data in this paper is previous survey evidence showing that economists agree more than non-economists on the determinants of health and health care expenditures (Fuchs *et al.*, 1998). Perhaps the most compelling reason for studying economists is that their beliefs about statistical and medical concepts can (in theory) be measured with far less noise than in the general population, whose poor understanding of statistics and "health literacy" is well documented (Williams *et al.*, 1995; Parker *et al.*, 1995; Baker *et al.*, 1998; Lusardi and Mitchell, 2009).

4.9 Conclusion

Economists are presumably as familiar with the normative benchmarks of consistency and accuracy as anyone. Yet they vary substantially in: (1) the degree to which their subjective beliefs adhere to the consistency requirements of probabilistic logic, (2) the accuracy of their beliefs, and (3) the PSA decisions (and other medical decisions) they make. Despite this variation, no positive associations between inconsistency and inaccuracy were observed. The data support the view articulated in Gilboa *et al.* (2009):

> We reject the view that rationality is a clear-cut, binary notion that can be defined by a simple set of rules or axioms. There are various ingredients to rational choice. Some are of internal coherence, as captured by Savage's axioms. Others have to do with external coherence with data and scientific reasoning. The question we should ask is not whether a particular decision is rational or not, but rather, whether a particular decision is more rational than another. And we should be prepared to have conflicts between the different demands of rationality. When such conflicts arise, compromises are called for. Sometimes we may relax our demands of internal consistency; at other times we may lower our standards of justifications for choices. But the quest for a single set of rules that will universally define the rational choice is misguided.

The conclusions we draw are not categorically in conflict with the possibility of real-world benefits from adhering to Bayes' Rule or other axioms based on internal consistency. If within-person divergence among plural normative criteria is typical, then our personal view is that consideration of these multiple normative criteria should be required to make meaningful normative comparisons between individuals and across different decision-making environments (*c.f.*, Berg, 2003, 2014). There seems to be a disconnect in the vast empirical literature on non-Bayesian beliefs by which Bayesian consistency is used to rank the rationality of individuals' beliefs without confirming whether Bayesian consistency matches the reward structure in which people apply their non-Bayesian beliefs. Why should we care about non-Bayesian beliefs in decision problems where consistency is not rewarded and there is no obvious mechanism guaranteeing that Bayesian beliefs tend to be more accurate?

References

Andriole, G. L., E. D. Crawford, R. Grubb, S. S. Buys, D. Chia, and T. R. Church. 2009. "Mortality Results From A Randomized Prostate-Cancer Screening Trial". *N Engl J Med.* 360(13): 1310–1319.

Aragones, E., I. Gilboa, A. Postlewaite, and D. Schmeidler. 2005. "Fact-Free Learning". *American Economic Review*. 95: 1355–1368.

Baker, D. W., R. M. Parker, M. V. Williams, and W. S. Clark. 1998. "Health Literacy and the Risk of Hospital Admission". *The Journal of General Internal Medicine*. 13: 791–798.

Bardsley, N., R. Cubitt, G. Loomes, P. Moffatt, C. Starmer, and R. Sugden. 2010. *Experimental Economics: Rethinking the Rules*. Princeton: Princeton University Press.

Barry, M. J. 2006. "The PSA Conundrum". *Archives of Internal Medicine*. 166: 38–43.

Baucells, M., J. A. Carrasco, and R. M. Hogarth. 2008. "Cumulative Dominance and Heuristic Performance in Binary Multi-Attribute Choice". *Operations Research*. 56(5): 1289–1304.

Behrens, J., W. Güth, H. Kliemt, and V. Levati. 2005. "Games that Doctors Play Two-layered Agency Problems in a Medical System". In: *Dimensions of Public Economics: Strategies and Policies Facing Uncertainty and Market Failure*. Ed. by A. Mehler and U. Cantner.

Berg, N. 2003. "Normative Behavioral Economics". *Journal of Socio-Economics*. 32: 411–427.

Berg, N. 2004. "No-Decision Classification: An Alternative To Testing for Statistical Significance". *Journal of Socio-Economics*. 35(8): 631–650.

Berg, N. 2014. "The Consistency and Ecological Rationality Approaches to Normative Bounded Rationality". *Journal of Economic Methodology*. 21(4): 375–395.

Berg, N. and U. Hoffrage. 2008. "Rational Ignoring with Unbounded Cognitive Capacity". *Journal of Economics Psychology*. 29(6): 792–809.

Blendon, R. J., J. M. Benson, M. Brodie, R. Morin, D. E. Altman, D. Gitterman, M. Brossard, and M. James. 1997. "Bridging the Gap between the Public's and Economists' Views of the Economy". *Journal of Economic Perspectives*. 11(3): 105–118.

Braga, J., S. J. Humphrey, and C. Starmer. 2009. "Market Experience Eliminates Some Anomalies-and Creates New Ones". *European Economic Review*. 53: 401–416.

Camerer, C. F. 1987. "Do Biases in Probability Judgment Matter in Markets?: Experimental Evidence". *American Economic Review*. 77: 981–997.

Camerer, C. F. and R. M. Hogarth. 1999. "The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework". *Journal of Risk and Uncertainty*. 19: 7–42.

Camerer, C., S. Issacharoff, G. Loewenstein, T. O'Donoghue, and M. Rabin. 2003. "Regulation For Conservatives: Behavioral Economics and the Case for "Asymmetric Paternalism"". *University of Pennsylvania Law Review*. 1151: 1211–1254.

Caplan, B. 2001. “What Makes People Think Like Economists? Evidence on Economic Cognition from the ‘Survey of Americans and Economists on the Economy’.” *Journal of Law and Economics.* 44(2): 395–426.

Caplan, B. 2002. “Systematically Biased Beliefs About Economics: Robust Evidence of Judgmental Anomalies from the Survey of Americans and Economists on the Economy”. *The Economic Journal.* 112(479): 433–458.

Carter, J. R. and M. D. Irons. 1991. “Are Economists Different, and If So, Why?” *Journal of Economic Perspectives.* 5(2): 171–177.

Ciatto, S. 2003. “Reliability of PSA Testing Remains Unclear”. *British Medical Journal.* 327: 750.

Ciatto, S., M. Zappa, R. Bonardi, and G. Gervasi. 2000. “Prostate Cancer Screening: The Problem of Overdiagnosis and Lessons to Be Learned from Breast Cancer Screening”. *European Journal of Cancer.* 36: 1347–1350.

Concato, J., C. K. Wells, R. I. Horwitz, D. Penson, G. Fincke, D. R. Berlowitz, G. Froehlich, D. Blake, M. A. Vickers, G. A. Gehr, N. H. Raheb, G. Sullivan, and P. Peduzzi. 2006. “The Effectiveness of Screening for Prostate Cancer: A Nested Case-Control Study”. *Archives of Internal Medicine.* 166: 38–43.

Cubitt, R. and R. Sugden. 2001. “Dynamic Decision-Making Under Uncertainty: An Experimental Investigation of Choices Between Accumulator Gambles”. *Journal of Risk and Uncertainty.* 22(1): 103–128.

Draisma, G., R. Boer, S. J. Otto, I. W. van der Cruijsen, R. A. M. Damhuis, F. H. Schröder, and H. J. de Koning. 2003. “Lead Times and Overdetection Due to Prostate-Specific Antigen Screening: Estimates from the European Randomized Study of Screening for Prostate Cancer”. *Journal of the National Cancer Institute.* 95(12): 868–878.

Emerson, R. W. 1841. *Selected Writings of Ralph Waldo Emerson.* New York, NY: Penguin.

Frank, R. H., T. D. Gilovich, and D. T. Regan. 1993. “Does Studying Economics Inhibit Cooperation?” *Journal of Economic Perspectives.* 7(2): 159–171.

Fuchs, V. R., A. B. Krueger, and J. M. Poterba. 1998. “Economists’ Views About Parameters, Values, and Policies: Survey Results in Labor and Public Economics”. *Journal of Economic Literature.* 36(3): 1387–1425.

Gaechter, S., H. Orzen, E. Renner, and C. Starmer. 2009. “Are Experimental Economists Prone To Framing Effects? A Natural Field Experiment”. *Journal of Economic Behavior and Organization.* 70: 443–446.

Ganguly, A., J. H. Kagel, and D. V. Moser. 2000. “Do Asset Market Prices Reflect Traders’ Judgment Biases?” *Journal of Risk and Uncertainty.* 20(3): 219–245.

Gann, P. H. 1997. “Interpreting Recent Trends in Prostate Cancer Incidence and Mortality”. *Epidemiology.* 8: 117–120.

Gann, P. H., C. H. Hennekens, and M. J. Stampfer. 1995. “A Prospective Evaluation of Plasma Prostate-Specific Antigen for Detection of Prostate Cancer”. *Journal of the American Medical Association.* 273: 289–294.

Gigerenzer, G. 2004. "Fast and Frugal Heuristics: The Tools of Bounded Rationality". In: *Blackwell Handbook of Judgment and Decision Making*. Ed. by D. J. Koehler and N. Harvey. Oxford: Blackwell. 62–88.

Gigerenzer, G., W. Gaissmaier, E. Kurz-Milcke, L. M. Schwartz, and S. Woloshin. 2007. "Helping doctors and patients make sense of health statistics". *Psychological Science in the Public Interest*. 8(2): 53–96.

Gilboa, I. 2010. "Questions in Decision Theory". *Annual Reviews in Economics*. 2: 1–19.

Gilboa, I., A. Postlewaite, and D. Schmeidler. 2008. "Probabilities in Economic Modeling". *Journal of Economic Perspectives*. 22: 173–188.

Gilboa, I., A. Postlewaite, and D. Schmeidler. 2009. "Is it Always Rational to Satisfy Savage's Axioms?" *Economics and Philosophy*. 25: 285–296.

Gilboa, I., L. Samuelson, and D. Schmeidler. 2010. "Dynamics of Inductive Inference in a Unified Model". *Working Paper*. Yale University.

Gilboa, I. and D. Schmeidler. 1995. "Case-Based Decision Theory". *Quarterly Journal of Economics*. 110: 605–639.

Gintis, H. forthcoming. "Rationality and its Discontents". *Economic Journal*.

Goldstein, D. G. and G. Gigerenzer. 2009. "Fast and Frugal Forecasting". *International Journal of Forecasting*. 25: 760–772.

Grunwald, P. D. and J. Y. Halpern. 2004. "When Ignorance is Bliss". *ACM International Conference Proceeding Series*. 70: 226–234.

Hammond, K. R. 1996. *Human Judgment and Social Policy: Irreducible Uncertainty, Inevitable Error, Unavoidable Injustice*. New York: Oxford University Press.

Harris, R. and K. N. Lohr. 2002. "Screening for Prostate Cancer: An Update of the Evidence for the U.S. Preventative Services Taskforce". *Annals of Internal Medicine*. 37(11): 917–929.

Hastie, R. and K. A. Rasinski. 1988. "The Concept of accuracy in Social Judgment". In: *The Social Psychology of Knowledge*. Ed. by D. Bar-Tal and A. W. Kruglanski. New York, NY; Paris, France: Cambridge University Press. 193–208.

Hawkes, N. 2006. "Prostate Cancer Test May Leave Men Even Worse Off". URL: http://www.timesonline.co.uk/tol/news/uk/article786762.ece.

Hogarth, R. M. and N. Karelaia. 2005. "Simple Models For Multi-Attribute Choice with Many Alternatives: When it Does and Does Not Pay to Face Tradeoffs with Binary Attributes". *Management Science*. 51(12): 1860–1872.

Hogarth, R. M. and N. Karelaia. 2006. "'Take-The-Best' And Other Simple Strategies: Why And When They Work 'Well' With Binary Cues". *Theory and Decision*. 61: 205–249.

Holmberg, L., A. Bill-Axelson, F. Helgesen, J. O. Salo, P. Folmerz, M. Haggman, S.-O. Andersson, A. Spangberg, C. Busch, S. Nordling, J. Palmgren, H.-O. Adami, J.-E. Johansson, and B. J. Norlén. 2002. "A Randomized Trial Comparing Radical Prostatectomy with Watchful Waiting in Early Prostate Cancer". *The New England Journal of Medicine*. 347(11): 781–789.

Kluger, B. D. and S. Wyatt. 2003. "Are Judgment Errors Reflected in Market Prices and Allocations? Experimental Evidence Based on a Monty Hall Problem". *Journal of Finance*. 59(3): 969–997.

Litwin, M. S., R. D. Hays, A. Fink, P. A. Ganz, B. Leake, G. E. Leach, and R. H. Brook. 1995. "Quality-of-Life Outcomes in Men Treated for Localized Prostate Cancer". *The Journal of the American Medical Association*. 273: 129–135.

Loewenstein, G. 1999. "Experimental Economics from the Vantage-Point of Behavioural Economics". *Economic Journal*. 109(453): 23–34.

Loewenstein, G. 2005. "Hot-Cold Empathy Gaps and Medical Decision Making". *Health Psychology*. 24: S49–S56.

Loewenstein, G. 2006. "The Pleasures and Pains of Information". *Science*. 312: 704–706.

Lusardi, A. and O. S. Mitchell. 2009. "How Ordinary Consumers Make Complex Economic Decisions: Financial Literacy and Retirement Readiness". *Working Paper*. Dartmouth University.

Parker, R. M., D. W. Baker, M. V. Williams, and J. R. Nurss. 1995. "The Test of Functional Health Literacy in Adults: A New Instrument for Measuring Patients' Literacy Skills". *Journal of General Internal Medicine*. 10: 537–541.

Robson, A. and L. Samuelson. 2009. "The Evolution Of Time Preference With Aggregate Uncertainty". *American Economic Review*. 99: 1925–1953.

Samuelson, L. 2001. "Analogies, Adaptation, and Anomalies". *Journal of Economic Theory*. 97: 320–367.

Savage, L. 1954. *The Foundations of Statistics*. New York, NY: Wiley.

Selten, R. 2001. "What is Bounded Rationality?" In: *Bounded Rationality: The Adaptive Toolbox*. Ed. by G. Gigerenzer and R. Selten. Cambridge, MA: MIT Press. 13–36.

Sorum, P. C., E. Mullet, J. Shim, S. Bonnin-Scaon, G. Chasseigne, and J. Cogneau. 2004. "Avoidance of Anticipated Regret: The Ordering of Prostate-Specific Antigen Tests". *Medical Decision Making*. 24: 149–159.

Spiegel, Y., A. Heifetza, and C. Shannon. 2007. "What To Maximize If You Must". *Journal of Economic Theory*. 133: 31–57.

Stanford, J. L., R. A. Stephenson, L. M. Coyle, J. Cerhan, Correa, J. W. Eley, F. Gilliland, B. Hankey, L. N. Kolonel, C. Kosary, R. Ross, R. Severson, and D. West. 1999. *Prostate Cancer Trends 1973–1995*. NIH Pub. No. 99–4543. Bethesda, MD: National Cancer Institute SEER Program.

Starmer, C. 2000. "Developments in Non-Expected Utility Theory: The Hunt for a Descriptive Theory of Choice under Risk". *Journal of Economic Literature*. 38: 332–382.

Starmer, C. 2005. "Normative Notions In Descriptive Dialogues". *Journal of Economic Methodology*. 12: 277–290.

Starmer, C. 2009. "Friedman's Risky Methodology". In: *The Methodology of Positive Economics: Reflections on the Milton Friedman Legacy*. Ed. by U. Mäki. Cambridge: Cambridge University Press. 285–302.

Steineck, G., F. Helgesen, J. Adolfsson, P. W. Dickman, J.-E. Johansson, B. J. Norlén, and L. Holmberg. 2002. "Quality of Life after Radical Prostatectomy or Watchful Waiting". *The New England Journal of Medicine*. 347(11): 790–796.

Studdert, D. M., M. M. Mello, W. M. Sage, C. M. DesRoches, J. Peugh, K. Zapert, and T. A. Brennan. 2005. "Defensive Medicine Among High-Risk Specialist Physicians In A Volatile Malpractice Environment". 293: 2609–2617.

Sugden, R. 1991. "Rational Choice: A Survey Of Contributions From Economics And Philosophy". *Economic Journal*. 101: 751–785.

Tversky, A. and D. Kahneman. 1974. "Judgment Under Uncertainty: Heuristics and Biases". *Science*. 185: 1124–1131.

U.S. Preventive Services Task Force. 2002. "Screening for Prostate Cancer: Recommendation and Rationale". *Annals of Internal Medicine*. 137(11): 915–916.

Wang, L. and K. Arnold. 2002. "Prostate Cancer Incidence Trends Reveal Extent of Screening-Related Overdiagnosis". *Journal of the National Cancer Institute*. 94(13): 957.

Wegwarth, O. and G. Gigerenzer. 2013. "Trust-Your-Doctor: A Simple Heuristic in Need of a Proper Social Environment". In: *The ABC Research Group, Simple Heuristics in a Social World*. Ed. by R. Hertwig and U. Hoffrage. New York: Oxford University Press. 67–102.

Williams, M. V., R. M. Parker, D. W. Baker, N. S. Parikh, K. Pitkin, W. C. Coates, and J. R. Nurss. 1995. "Inadequate Functional Health Literacy among Patients at Two Public Hospitals". *The Journal of the American Medical Association*. 274(21): 1677–1682.

Wisdom, J., J. S. Downs, and G. Loewenstein. 2010. "Promoting Healthy Choices: Information Versus Convenience". *American Economic Journal: Applied Economics*. 2: 164–178.

Yao, S.-L. and G. Lu-Yao. 2002. "Understanding and Appreciating Overdiagnosis in the PSA Era". *Journal of the National Cancer Institute*. 94(13): 958–960.

Yezer, A. M., R. S. Goldfarb, and P. J. Poppen. 1996. "Does Studying Economics Discourage Cooperation?: Watch What We Do, Not What We Say or How We Play". *Journal of Economic Perspectives*. 10(1): 171–186.