Who's to Blame for Survey Instability: Respondents with Nonexistent Preferences or Researchers with Flawed Measures?*

Libby Jenke† Gary King‡

June 17, 2026

Abstract

Neither. For at least 75 years, survey researchers have found that about 25% of respondents give different answers when asked the same question twice (even if no material changes occur and respondents do not remember being asked the first time). This "survey instability" problem casts doubt on a vast research enterprise spanning large areas of academia and industry, is core to many ongoing substantive debates, and requires a resolution for proper survey design and analysis methods. We collect a wide variety of observational and experimental evidence, including 59 unique surveys. We first show that instability barely drops after accounting for both existing explanations, i.e., when respondents have fixed knowledge of their preferences and researchers use high quality, unbiased survey instruments. We trace a large component of survey instability to a different source recognized only in fields with non-survey measurement instruments — intrinsic human stochasticity. We then trace the decision making, cognitive, psychological, and individual characteristic precursors of this stochasticity and reveal their wide ranging implications for understanding respondents, avoiding inattention, designing surveys, and building statistical analysis methods.

*Our thanks to Dominic Skinnion and Katalina Toth for superb research assistance; Mike Alvarez, Soubhik Barari, Matt Blackwell, Francesco Bilotta, Randy Buckner, Danny Ebanks, Dan Gilbert, Chase Harrison, Kosuke Imai, Jonathan Katz, Ella King, Gabe Lenz, Caroline Martin, Tom Palfrey, Antonio Rangel, Kota Saito, Brandon Stewart, and Dan Schacter for helpful suggestions; and the Roper Center for its collection of survey data and questions. See GaryKing.org/instability for the current version of this paper and its Supplementary Appendix.

†Assistant Professor, Department of Political Science, University of Houston; LibbyJenke.com, LJenke@uh.edu.

‡Albert J. Weatherhead III University Professor, Institute for Quantitative Social Science, Harvard University; GaryKing.org, King@Harvard.edu.

1 Introduction

Why do ∼25% of survey respondents answer identical binary choice questions in different ways when given the opportunity at separate times, even if nothing material has changed and they do not remember being asked before? The source of this widespread *survey instability* has been an open question for at least three-quarters of a century (Lazarsfeld, 1948) and is at the core of many central substantive questions and methodological choices.

In political science, where survey research is used as evidence in half of all quantitative articles (King, Honaker, et al., 2001; Sturgis and Luff, 2020), the puzzle of survey instability sustains a longstanding debate about the meaning of American democracy: Do Americans have ephemeral "nonattitudes" (making democracy and learning from survey data impossible) or does measurement error in survey instruments obscure stable political and policy knowledge (Converse, 1964)? Whether survey instability is the fault of citizens, with nonexistent preferences, or researchers, with flawed survey instruments, is also central to debates in numerous other scholarly areas, such as religiosity (Lim, MacGregor, and Putnam, 2010), immigration attitudes (Kustov, Laaker, and Reller, 2021), time discounting in health behavior (Kang and Ikeda, 2014), emotions (Lee, Amir, and Ariely, 2009), future-oriented attitudes (Preuss, 2021), consumer choices (Lee, Lee, et al., 2015), and even musical preferences (Rentfrow, Goldberg, and Levitin, 2011).

For methodological purposes also, studying survey instability (or "reliability") can teach us about the data generation process for survey data, on which survey design and statistical analysis methods depend. Different data generation processes imply different design and analysis methods and sometimes diametrically opposing empirical results. (For one example, unlike for a continuous outcome variable in a linear regression, random error in typically discrete survey outcomes biases substantive conclusions (e.g. Westwood et al., 2022).) An explanation for survey instability would have implications for both our deep substantive understanding of the people we study and practical applications across the entire survey research enterprise, spanning many areas of academia and industry.

We begin with a precise statement of the survey instability problem and the methods and research strategy we use to tackle it, along with baseline estimates (Section 2). We

then summarize explanations in the literature, which provide reasons to question the sufficiency of either or both existing explanations, along with additional empirical tests (Section 3). We outline we we propose as the main cause of survey instability that has not previously been analyzed, "intrinsic human stochasticity" (Section 4), and reveal its causes, at levels from most to least proximate, including decision making, cognitive processes, psychological states, and individual characteristics (Section 5). We offer highly diverse empirical evidence, including 59 surveys covering repeated conjoint and traditional survey questions; ANES wording experiments; priming experiments; fixed belief decision studies; time-on-task observations and experiments; web-based eye tracking; mind-wandering probes; preoccupation, attention check, persona, and mind-wandering instability predictions; bot detection; and age heterogeneity studies. We conclude by showing how these findings generate specific, practical advice for the design and implementation of surveys (Section 6).

2 Problem and Methods

We now describe the problem to be solved and how we plan to solve it. This includes a precise definition of survey instability, along with the necessary assumptions (Section 2.1); a description of an experimental protocol we designed to satisfy these assumptions (Section 2.2); baseline estimates of survey instability (Section 2.3); and a new strategy we follow for conducting and learning from surveys made possible by recent changes in survey technology (Section 2.4).

2.1 Defining and Estimating Survey Instability

Consider a population of N potential respondents and, for simplicity but without loss of generality, a binary survey question. Let $C_{it} \in \{0, 1\}$ denote the choices respondent i ($i = 1, \ldots, N$) would give to this question if asked at times $t = 1$ and $t = 2$. Survey instability for individual i is $D_i = \mathbf{1}(C_{i1} \neq C_{i2})$ (which equals 1 if $C_{i1} \neq C_{i2}$ and 0 otherwise). Population average survey instability is the proportion $\Delta = \frac{1}{N} \sum_{i=1}^{N} D_i$. Finally, for a sample of $n < N$ respondents randomly selected from the population, the

sample average survey instability we observe is $D = \frac{1}{n}\sum_{i=1}^{N} S_i D_i$, where S_i is 1 if i is sampled and 0 otherwise ($1 - D$ is also known as "intra-respondent reliability").

To give meaning to sample D and population Δ survey instability, we add three assumptions only implicitly addressed in the literature. (Without meeting all three assumptions, D has a different meaning than typically intended or, put differently, is a biased estimate of Δ.)

First, *no material change* has occurred between the two time points that could provide a reason for instability. Most violations of this assumption would include information received by the respondent due to changes in the world or the survey instrument that changes some aspect of the meaning of the question. Second, at $t = 2$, the respondent has *no memory* of being asked the question or what their answer was at $t = 1$. Empirically, these two assumptions tend to work against each other: If the two survey questions are well separated in time, the "no memory" condition is more likely to be satisfied but the "no material change" condition is less likely. In contrast, if the two questions are asked near each other, the reverse is true. The literature has dealt with this trade off by asking the two questions in separate waves of panel surveys weeks, months, or years apart, but this leads to yet another problem: the potential biasing effects of sample attrition. (If, for example, respondents who remain in the survey for both waves are more stable than those who leave, then D will be a downwardly biased estimate of Δ (Backström et al., 2025)). We thus clarify by adding a third and final assumption: *attrition at random*, meaning that attrition is unrelated to the variable of interest, which in our case is survey instability.

Survey instability has a logical range of $\Delta \in [0, 1]$, but under these three assumptions, the practical maximum is 0.5, indicating choices are based on the equivalent of coin flips (expected instability under independent Bernoulli draws).

2.2 Experimental Setup

For expository simplicity, we summarize each survey question with a code composed of an uppercase letter and an integer, sometimes followed by a lowercase letter (e.g., Q1a). The uppercase letter indicates the question type (Q = conjoint question, B = conjoint burn in, D = conjoint distractor, M = mind wandering probe, P = preoccupation, S =

socioeconomics or demographics, V = vote, T = randomized treatment, or others). The integer distinguishes different questions of the same type (1, 2, ...), and the optional lower case letter distinguishes among repeated identical questions (a, b, ...). Collections of codes, such as B1 Q1a D1 D2 D3 Q1b, summarize a single treatment regime within a single survey instrument (see Supplementary Appendix A9 for details).

Our setup builds on Clayton et al. (2025), which compares answers to two identical forced-choice binary conjoint questions administered to the same respondents in the same survey, separated by several conjoint distractors (randomized from the same attribute list, as is standard for conjoints), such as Q1a D1 D2 D3 Q1b. Because Q1a and Q1b are asked only moments apart, the "no material change" assumption is almost surely satisfied. Perhaps surprisingly, the "no memory" assumption also appears to be satisfied: Clayton et al. (2025) administered surveys to 9,472 respondents, and not one mentioned seeing a repeated question. We also conducted other tests of the no memory condition, leading to the same conclusion (see Supplementary Appendix A2). Finally, because Q1a and Q1b are in the same survey, the "attrition at random" assumption is routinely satisfied by avoiding almost all attrition.

We tried many other variants of this setup, including different numbers of distractor questions, burn-in questions, different types of filters, and various combinations of these and other modifications. None made a material difference to instability estimates (see Surveys 1 and 2). In addition, although the complicated nature of the conjoint question is a valuable aspect of this experimental platform, we show in Section 5.2 that many traditional, non-conjoint survey questions can often also be asked in the same survey too without violating our assumptions, so long as sufficient distractors are included. In our experiments, we use conjoints when feasible and other questions when useful or necessary.

2.3 Baseline Estimates of Survey Instability

We compute two sets of survey instability baselines spanning the history of survey research. Despite being separated by three-quarters of a century, the estimates are indistinguishable.

First, we obtain the earliest panel survey with sufficient documentation available, the 1948 Elmira Study (Lazarsfeld, Berelson, and McPhee, 1948, codebook, p.296). We then estimate survey instability in the same reported vote choice in two surveys a month apart. To avoid an "attrition at random" assumption, we switched from a point estimate to a bound. The results indicate survey instability in this survey falls in the interval $D \in [0.11, 0.33]$, with a midpoint of $D = 0.22$.[1]

Second, we estimated instability in 53 (of our 59) surveys, using different versions of our experimental protocol (such as B1 Q1a D1 D2 D3 Q1b; see Section 2.2), and present them all in Figure 1. We introduce details below but for now note that they are highly diverse, conducted over three years on different online survey platforms, in different formats, and each with different treatment regimes to test different hypotheses. All are nontrivial binary questions of interest to social scientists, such as about policy or candidates, usually in standard conjoint format (extensions to other question types, numbers of response categories, and priming appear below too).

For each survey, we provide one estimate of survey instability (leaving estimates from different treatment regimes for later). The figure shows that, on average across all our surveys, a remarkable 23.0% of respondents give different answers to two identically worded binary questions. This figure is nearly *half* of what we would expect if respondents were flipping coins. Variability in instability across surveys is small, with a few outliers explained by the particular type of experiment we were running. The vast majority of the surveys fall within the range of the 1948 estimates, the midpoint of which (0.22) is almost the same as the mean estimate (0.23) here.

2.4 Research Strategy

Both existing explanations for survey instability (Section 3), as well as our proposed alternative (Section 4), are broad conceptual theories or frameworks. We thus follow the literature's standard for testing these frameworks — identifying some of their "observable

[1] We also collected as many reported instability estimates as we could find in the scholarly literature; although the survey questions, measures, and validity of the three assumptions varied greatly (and so they do not admit to a systematic summary), we could detect no systematic change in the magnitude of these estimates.

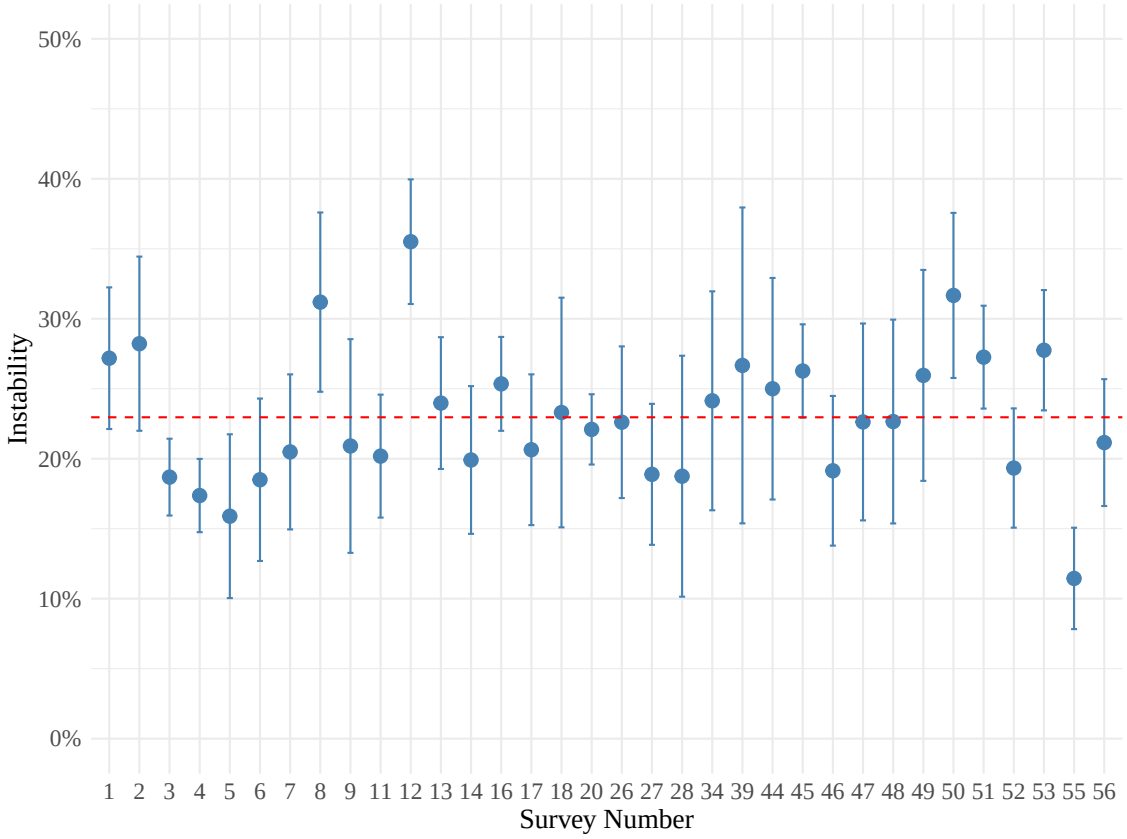


Figure 1: Survey Instability in 53 Diverse Surveys over Three Years, average = 23.0%. (See Supplementary Appendix A9 for details of each survey summarized here.)

implications," collecting observations on each, and determining whether they are consistent with theoretical predictions (see King, Keohane, and Verba, 1994, Section 1.3.1). These frameworks should not to be confused with narrowly defined treatments used in causal inference (which we use when they are observable implications). The open ended nature of broad theories like these always leaves other observable implications to future research, and so their veracity is best judged by their ability to predict the outcome of implications that have not yet been observed.

For specific survey tests, we improve on standard practices in five ways, all of which take advantage of the recent rise of online survey platforms (such as Lucid, Mechanical Turk, Prolific, etc.). For nearly all of prior survey history, experimenting with new survey questions required either lobbying organizers of large omnibus surveys or fielding a full survey on your own. Sophisticated studies of survey instability then involved building (necessarily) assumption-laden models of the relatively rare multi-wave panel studies

(e.g., Brady, 1993; Erikson, 1979; Hout and Hastings, 2016). In contrast, this new survey technology (1) cuts preparation time from months to hours, (2) reduces costs, and (3) enables researchers to avoid some ex post modeling assumptions via ex ante survey design choices.

More fundamentally, the core of any research strategy involves making oneself vulnerable to being proven wrong and learning from the process. Since using the same data for hypothesis generation and validation removes such vulnerability, best practice under the previous survey technology regime involved pre-registering hypotheses to avoid just-so stories or p-hacking. In contrast, (4) under the new survey technology, researchers can easily run many out-of-sample tests in fresh data, each producing additional chances to correct mistaken hypotheses and develop better theory. Thus, instead of offering one set of speculative hypotheses followed by a single test (or preregistration which, in this situation, would be performative), we adjust initial hypotheses as we learn from many sequentially run experiments, all with reduced risk of p-hacking. We make ourselves vulnerable to being proven wrong by repeatedly rerunning surveys from scratch in newly collected data and encourage others to replicate our results in new data. In practice, we continued modifying our ideas and running surveys to test them until we were able to accurately predict the outcome of tests of our ideas in new survey data. This strategy provides more chances to be wrong and thus the ability to progress faster, learn more, and — by bringing more data to bear on the problem — generate more confident conclusions.

Finally, (5) we take advantage of this technology to offer three types of research transparency, ordered from customary to novel. First, following now standard practice, we make publicly available in Dataverse a replication archive with all the data and code necessary to reproduce our statistical results, beginning with the data we collected (King, 1995, 2007). Second, we provide easy ways of replicating our entire data collection processes from scratch in new data and new populations. And third, we try to make transparent where we were not merely vulnerable to being proven wrong, but were actually wrong along the way to our final conclusions. This includes many tests we conducted where our early hypotheses were incorrect, survey designs were flawed, question wordings were

confusing, experimental setups were defective, attempts at survey measurement had unintended treatment effects, and overall theories of survey instability were inadequate to our goals (see Supplementary Appendix A9).

Online survey platforms do not solve all problems with the existing survey regime. To deal with the power of the platform, we also developed some proposed ethical standards for their use (see Supplementary Appendix A8). Additionally, most platforms promise some form of "representativeness" but neither they nor traditional surveys are able to overcome the epidemic of nonresponse to guarantee true probability samples. Survey researchers have also long been aware that the survey mode can in some cases impact survey responses. Finally, we are able to provide evidence that severe contamination by AI bots is unlikely (see Figure 6 for details).

3 Prior Explanations

The literature offers two conflicting explanations for survey instability: respondent preferences on major issues are *nonexistent* or real but *observed with noise*. The first is a claim about the nature of human beings who have no actual preferences over major issues of the day, and thus respond whimsically or randomly, without any anchor to genuine views. The second is interpreted as due to measurement error in the survey instrument.

Nonexistent Preferences The first and most extreme position in the literature is that of Converse (1964) who argued that high levels of survey instability on major issues commonly debated in presidential election campaigns stems from respondents with nonexistent preferences (what he called "nonattitudes"). This may include respondents who do not understand the subject of the question or the response options. While we all have no preferences on or knowledge about some issues, subsequent social scientists have repeatedly shown that ordinary citizens have genuine preferences on a wide range of topics and rarely change them without a reason. Perhaps the most compelling reason to reject the nonexistent preferences explanation is the vast array of systematic, replicated findings from survey-based scholarship on attitudes, opinions, and behaviors in thousands of aca-

demic articles across fields over three-quarters of a century. This point has been formally demonstrated via model-based methods (e.g., Achen, 1975; Erikson, 1979) and modern model-free approaches (Ansolabehere, Rodden, and Snyder, 2008).

Nonexistent preferences on major issues would be particularly surprising given that even intentional attempts to change reported preferences usually fail because of a range of psychological phenomena including "epistemic vigilance" (aka, good BS detectors; Sperber et al. 2010), "psychological immune systems" (Gilbert, Pinel, et al., 1998), "status quo bias" (Samuelson and Zeckhauser, 1988; Godefroid, Plattfaut, and Niehaves, 2023), "belief perseverance" (Wilson and Brekke, 1994, p.122), confirmation or "myside" bias (Mercier, 2022), and self-consistency instincts (Isenberg and Brauer, 2022). Indeed, even the most extensive attempts at persuasion throughout history have had highly limited effects, as political candidates in democracies, dictators of totalitarian regimes, and marketing directors for commercial products consistently find (Mercier, 2020). Changing behavior and attention is sometimes possible, but the high instability in most surveys on major issues is unlikely to be fully explained by nonexistent preferences.

Measurement Error The second explanation for survey instability is the popular idea that preferences are observed with noise, generated by measurement error due to our flawed survey instruments. Researchers have in fact shown through decades of hard work how to remove numerous forms of survey bias — including question order, wording, semantics, translation, interpersonal incomparability, interviewer or mode effects, sample designs, etc. (Blackwell et al., 2025; Hopkins and King, 2010; Berinsky, 2017; Schwarz, 2007b; Schwarz, 1999; Stantcheva, 2023; Vannette and Krosnick, 2017; Schaeffer and Dykema, 2020; King, Murray, et al., 2004). However, if a survey improvement reduces bias in Q1a and Q1b in the same way, its effect on instability, which is a function of their difference, could still be zero. For example, adding the word "baby" to a question about abortion (or to the prior question for priming) may dramatically increase the prevalence of pro-life survey responses, but we would not expect that adding the word to both questions would change instability since both would increase by roughly the same amount. Most survey improvements are effective at reducing bias, but they are not usually intended, ex-

pected, or found to reduce instability. Indeed, despite the substantial progress researchers have made in reducing survey instrument bias over many decades, we have been unable to detect any historical reduction in estimates of survey instability (see Section 2.3).

We now offer nine more direct experiments that test this hypothesis and which show, consistent with what we infer from the literature, that measurement error is not the primary cause of instability.

First, we study question wording improvements designed to bring the 2008 American National Election Survey (ANES) in line with the best survey practices. To do this, we first ran four randomized experiments comparing the original ANES questions to the new ones designed to reduce measurement error. The problems fixed in the individual questions include a lack of response labels (Surveys 40 and 41), a simplification from two questions to one shorter one (Survey 42), and a double barreled question (Survey 43), along with other smaller wording improvements. The vertical axis of Figure 2 indicates the average instability difference between those randomly assigned the original vs. changed question wordings (in the first four results at the left). Point estimates appear as dots with 50% and 95% confidence intervals, along with a horizontal line marking no difference. These four tests (shown at the left in the same order as the survey number), reveal a small increase in instability for two questions and a slightly larger decrease for the other two, but with none significantly different from zero. Taken together, these results suggest that these improved question wordings do not account for most survey instability.[2]

Finally, we study priming, the hypothesis that respondents hold multiple conflicting "considerations" and answer each survey question depending on which is activated at the time (Zaller and Feldman, 1992). According to this logic, the question immediately preceding Q1a and Q1b may prime the respondent to answer in different ways, thus producing instability. Of course, numerous tests confirm the impact of priming, but we sought to test whether this matters in our conjoint-based experimental platform. We thus randomly assigned respondents to a control group, which received our standard question battery

[2] Although measurement error does not seem to explain observed levels of instability, some types might still have an effect. We tried to show this by generating an exception that proves the rule, but we only managed to produce the predicted effect by wording a question far more confusing than any that we have seen used in practice. See Survey 36, with wording like "Which candidate do you not dislike the most?"

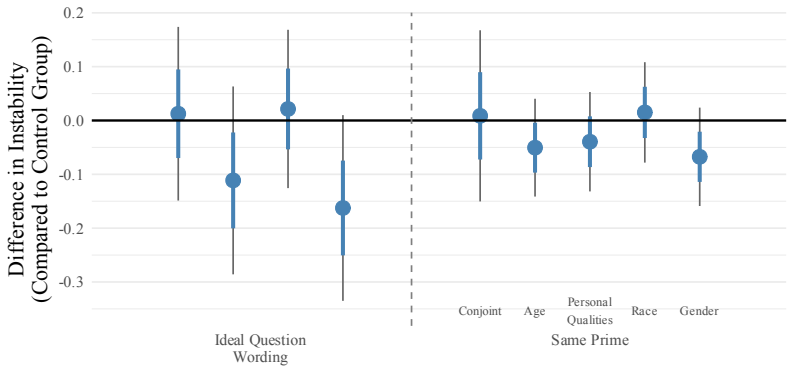


Figure 2: Reducing Bias Does Not Reduce Survey Instability. The effects of ANES question wording changes on instability (at the left corresponding to Surveys 40–43, respectively) and of the same questions before Q1a and Q1b to test priming (at the right; Surveys 44 and 58).

with a separate conjoint as a burn in and our usual randomly assigned attributes (B1 Q1a D2 D3 D4 Q1b), and the treatment group, which received repeated identical questions serving as a treatment, T4a and T4b, before both Q1a and Q1b (T4a Q1a D2 D3 T4b Q1b). We tried five different primes—a conjoint question (Survey 44) and primes suggesting that voters should prefer candidates with certain ages, personal qualities, races, and gender (Survey 58). If the considerations have a noticeably large effect on instability, the treatment group—consistently primed by the same content—should show more stable responses. However, we find a near zero instability difference (the result at the right of Figure 2), suggesting that priming by different considerations, even if important in reducing bias, is not a major contributor to survey instability.[3]

The evidence in all five experiments is consistent with that implied by the literature: survey measurement error does not account for large portions of the observed levels of survey instability.

4 Intrinsic Human Stochasticity

The literature in Section 3 logically partitions explanations for survey instability into respondents with preferences that are (1) nonexistent and (2) observed with noise, but with little positive evidence for either. We now enhance the second explanation, which has

[3] As shown in Figure A3, these questions were effective in priming bias on individual questions.

previously been interpreted as due to measurement error. The inadequacy of this view has long been suspected: as Zaller and Feldman (1992) points out, this explanation is "underspecified at its theoretical core. When, as all estimates agree, measurement 'error' typically constitutes one-half or more of the variance of typical attitude items, one naturally wonders what exactly this 'error' consists of and how it has been generated. Yet we presently know so little about these questions that the term remains essentially an alternative name for 'unexplained variance'." (Emphasis removed.)

We propose that conceptualizing "preferences observed with noise" as "measurement error" omits an important source of observed noise. In fact, in many areas devoted to measurement outside of survey research, researchers recognize this noise as having *two* sources: (a) *measurement error* and (b) *intrinsic stochasticity* (see Figure 3). We propose to extend this to survey research. This crucial distinction, between variability due to the measuring instrument and fundamental stochasticity of the object being measured, is referred to as technical vs. biological variation (in biology; Molloy et al. 2003; Bryant et al. 2011), estimation vs. fundamental uncertainty (in quantitative social science; see King 1989), thermal noise vs. physiological error (in fMRI studies; Triantafyllou et al. 2005; Brooks et al. 2013), the limits of optical microscopy such as diffraction and representation vs. Brownian motion (in soft matter physics; King, Engel, et al. 2025; Rose et al. 2020), measurement error vs. individual differences (in psychometrics; Hajcak, Meyer, and Kotov 2017), among others.

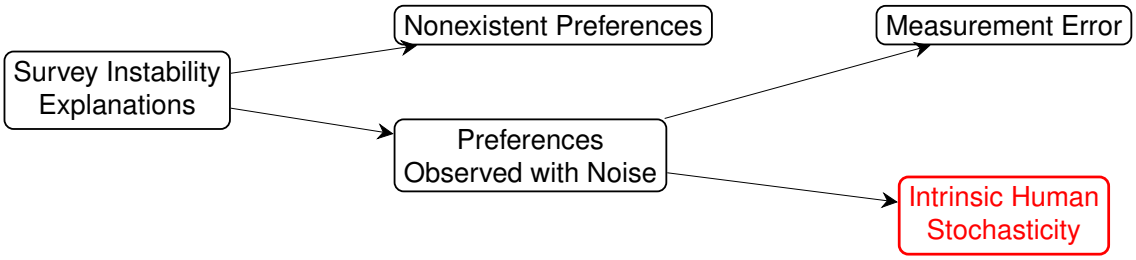


Figure 3: Explanations of Survey Instability, from the literature and our proposed explanation (in red).

To be clear, the "intrinsic" in "intrinsic stochasticity" does not refer to a component part of the subject of study; it means "essential to its existence". Because electrons have an intrinsic negative charge, the charge cannot be removed without destroying the electron.

Similarly, stochasticity is intrinsic to human beings and thus survey responses. A creature with zero stochasticity is not an "attentive respondent;" it is not even a (live) human being. Perhaps surprisingly, intrinsic stochasticity is both *omnipresent* and *beneficial* to the organism.

In the first, stochasticity is generated everywhere in biology — by failures of memory, fluctuations in other brain activities, environmental influences, neural noise and activation, path dependence in memory retrieval, subconscious desires or worries, variation in sensory noise, etc. Indeed, a "person" is not even a fixed entity over time, as 1.2% of every human's cells are replaced every day (Sender and Milo, 2021), as are many more of the bacteria and viral particles that live within us (which themselves vastly outnumber our own cells; Sender, Fuchs, and Milo 2016; Liang and Bushman 2021). (If you felt like a new person when you woke up this morning, you were right, literally.) For a particularly vivid example of the role of randomness, olfactory sensations have well-documented random representations in the brain's piriform cortex (Stettler and Axel, 2009; Sosulski et al., 2011), a feature conserved across species over evolutionary history (Caron et al., 2013).

In the second, instead of being a "bug," this omnipresent stochasticity turns out to be a "feature" that optimizes many specific biological processes (Heams, 2014; Ilan, 2020), at "all scales of brain activity" (Rolls and Deco, 2010; McDonnell, Goldwyn, and Lindner, 2016, p.5), and even, most prominently, Darwinianism. The advantage of just the right amount of randomness is known as "stochastic facilitation" and enhances learning, creativity, the stability of cognitive states, recall, decision making, information processing, the efficiency of neural networks, perception, and many other areas (McDonnell and Ward, 2011).[4]

[4]As an analogy, consider performing a Computed Tomography (CT) scan on the same patient twice in a row. Even with a perfect scanner, the scans would differ because of flowing blood, electrical and chemical signaling, a pumping heart, moving lungs and diaphragm, a churning stomach, synaptic activity, glands releasing hormones into the bloodstream, peristalsis moving food through the digestive tract, muscles twitching, eyes moving, food and hydration influencing organ size, metabolic rate changes, etc. Survey responses are like CT images: The same squishy, moving, flowing, signaling, squeezing, twitching creature in the scanner is the one we're asking questions of in a survey and must be present in what comes out of the same patient's mouth when asked a question. It just turns out that intrinsic stochasticity is part of the total survey error researchers try to reduce (see Groves and Lyberg, 2010). Hints from the literature about the presence of intrinsic human stochasticity can be seen in labels used for this extra variability, such as "random responding" (Pinsoneault, 2007), respondent "reliability" (Tourangeau, 2021; Alwin and Krosnick, 1991), "inattentive respondents," (Alvarez et al., 2019) "careless respondents" (Meade and Craig,

Intrinsic human stochasticity is a conceptual theory or framework (Section 2.4), just like previous explanations of survey instability (Section 3). We therefore test it in the same way — by finding observable implications, collecting data, and comparing them to their theoretical predictions.

5 Sources of Intrinsic Human Stochasticity

We now outline key sources of intrinsic human stochasticity, each providing additional observable implications of this explanation for survey instability. Because the "causes of an effect" are not as well defined as the "effects of a cause" (Holland, 1986), we need to narrow our goals (after all, the big bang, the birth of the survey respondent, and a finger muscle clicking an inconsistent answer are each full explanations of survey instability, but obviously not useful for social science research). Thus, in our search, we choose to focus on only those sources useful for the goals of social science research, which include (1) understanding the people social scientists study and (2) improving survey design and analysis.

Figure 4 offers a high level summary of our explanation of the sources of intrinsic human stochasticity, with survey instability at the right side and the sections below that follow, progressing backwards, from most proximate at the right to the least proximate at the left. These include survey response decision making (with evidence in Section 5.1), cognitive processes (Section 5.2), psychological states (Section 5.3) and individual characteristics (Section 5.4).

5.1 Decision Making

We now describe the most proximate of the causes of survey instability we study, the decision made when answering a survey question (See Figure 4), and identify observable implications of our intrinsic human stochasticity framework. Unfortunately, the vast literature on human decision making is not often used in the survey instability literature and so we mine the former to enhance the latter. In particular, we focus on the distinction be-

2012), "inconsistent responding" (McGrath et al., 2010), "non-content based responses" (Arias et al., 2023), and "insufficient effort respondents" (Huang et al., 2012).

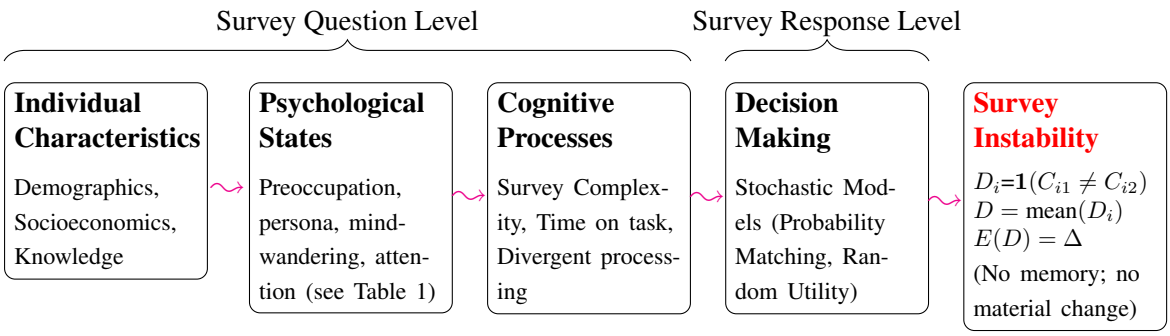


Figure 4: Sources of Intrinsic Human Variability as an Explanation for Survey Instability. (The symbol "⇝" means "leads to" and indicates a rough temporal ordering, not necessarily causality. Black boldface corresponds to sections that follow.)

tween *deterministic* and *stochastic* decision making — with only the latter a contributor to intrinsic human stochasticity that could generate, and help explain, survey instability. We first clarify results from this literature with a simpler mathematical representation, and then offer our own empirical experiments, all of which confirm that survey decision making has a large intrinsic stochastic component, even with minimal possibilities for impacts of measurement error, indifference, and respondent uncertainty about the decision environment.

Recall that the respondent's (observed) choice in answering a binary question is $C_{it} \in \{0, 1\}$ (see Section 2.1). We now also define a respondent's "preference" as a fixed but unobserved state indicating which option is better for them and denote it as $\rho_{it} \in \{0, 1\}$. Because respondents are uncertain about the value of ρ_i, they form beliefs about which option is better, which we formalize as π_i, the probability that $\rho_i = 1$. Under the "no material change" condition, we simplify these as $\rho_{i1} = \rho_{i2} \equiv \rho_i$ and $\pi_{i1} = \pi_{i2} \equiv \pi_i$.

Our goal is to show under what circumstances observed survey instability D_i is generated during repeated mappings from known fixed beliefs π_i (about unknown preferences ρ_i) to observed choices C_{it}.

Deterministic Models A deterministic model of decision making turns beliefs into choices without error: if a respondent's knowledge of their preferences remains the same, $\pi_{i1} = \pi_{i2}$, then their survey choices should remain identical $C_{i1} = C_{i2}$ so that $D_i = 0$.

Although numerous deterministic decision making models can be constructed (e.g.,

"always choose the option in the first column"), the rational choice *expected utility model* is the most common. In this model, respondents make survey choices that maximize their utility by maximizing the probability that they opt for their true preference:

$$C_{it} = \mathbf{1}(\pi_i > 0.5) \tag{1}$$

where the indicator function $\mathbf{1}(a) = 1$ if a is true and 0 otherwise. Although this model is certainly used by ordinary people in many situations, and is useful as a first step in modeling, "a fairly large body of experimental evidence... shows that subjects systematically make choices that violate properties required by expected utility" (Keller, 1992), especially in low-stakes situations like survey responses (Jenke et al., 2021).

Stochastic Models We partition stochastic decision models based on whether stochasticity is thought of as (a) an input, as most common in economics, or (b) a fundamental consequence of cognitive heuristics (or other biological variation), as is more common in psychology. The economics approach is more mathematically encompassing; the psychology approach may be more consistent with empirical evidence.

In economics, the *random utility model* generalizes the expected utility model in Equation 1 by adding a mean zero error term, η_{it}, generated just before the decision is made:

$$C_{it} = \mathbf{1}(\pi_i + \eta_{it} > 0.5) \tag{2}$$

Because $\pi_i + \eta_{it}$ is the best information the respondent has about their beliefs, who then decides rationally, this decision making process is known as "stochastic rationality".[5]

In psychology, the most popular stochastic decision making model is known as *probability matching*, which we write as

$$C_{it} \sim \text{Bernoulli}(\pi_i) \tag{3}$$

where choices are independent over t. For those accustomed to studying or assuming rationality, this model can be perplexing because the probability that the respondent makes

[5]Economists usually derive these models via reasonable but unnecessary stories about respondents maximizing unobservable utilities. Under Equation 1, the utility for option 1 is $U_{it1} = \pi_i(1) + (1-\pi_i)(0) = \pi_i$ and 0 is $U_{it0} = \pi_i(0) + (1-\pi_i)(1) = 1 - \pi_i$. Under Equation 2, errors ϵ_{it1} and ϵ_{it0} are generated just before each choice, and the resulting utilities are $U_{it1} = \pi_i + \epsilon_{it1}$ and $U_{it0} = 1 - \pi_i + \epsilon_{it0}$. To derive the simpler version of the model we give in Equation 2, let $\eta_{it} = (\epsilon_{it1} - \epsilon_{it0})/2$.

a choice equal to their preference is lower than under expected utility.[6] However, the evidence overwhelming supports it: "probability matching has been observed in thousands of geographically diverse human subjects over several decades, as well as in other animal species, including ants, bees, fish, pigeons, and primates [lengthy citations omitted]. In virtually any setting where a [human or nonhuman] animal is able to make a choice between A versus B in a randomized experiment, we observe probability matching" (Lo, Marlowe, and Zhang, 2021).

Interpretation Survey instability is obviously zero for respondents following the deterministic expected utility model (Equation 1). It is positive for those using the stochastic random utility (Equation 2) or probability matching (Equation 3) models. To show this for both stochastic models, let $\Pr(C_{it} = 1) \equiv p$, and then expected individual survey instability is

$$\Delta_i = p(1-p) + (1-p)p = 2p(1-p) > 0, \tag{4}$$

where $\Delta = \sum_{i=1}^{N} \Delta_i / N > 0$, and so survey instability is positive.[7]

Our simple reformulation of these models also makes clear their connections. First, as a mathematical entity, η_{it} in the random utility model is unobserved and so can be anything, but it is normally interpreted as producing uncertainty about knowledge of π_i, and humans as always making deterministic, rational decisions, conditional on all available information. This (and the simple math involved) implies that expected utility (Equation 1) is a special case of random utility (Equation 2) with $\eta_{it} = 0$. Second, while the random utility model is *deterministic given stochasticity, i.e.,* η_{it}, the probability matching model is *irreducibly stochastic*: If π_i is known to the respondent (and η_{it} is irrelevant or zero),

[6] For example, if (for any one survey question) $\pi = 0.7$ then, under the rational mechanism, they choose $C = 1$ every time, which gives them their (true) preference 70% of the time. In contrast, under probability matching, the respondent now only chooses $C = 1$ with probability $\pi = 0.7$ and, of those choices, they are right only 0.7 of the time; similarly, they choose $C = 0$ with probability $1 - \pi$ and are only correct $1 - \pi$ of the time. Thus, under probability matching, the respondent chooses their preference with probability $\pi^2 + (1-\pi)^2 = 0.58 < 0.7$.

[7] Under probability matching, Equation 4 specializes to $p = \pi_i$ and so $\Delta_i = 2\pi_i(1-\pi_i) > 0$. Under the random utility model, the expression specializes to $p = \Pr(C_{it} = 1) = \Pr(\eta_{it} > 0.5 - \pi_i) = 1 - F(0.5 - \pi_i)$, where $F(\cdot)$ is an assumed continuous cumulative distribution function of η_{it}, and so $\Delta_i = 2[1 - F(0.5 - \pi_i)] \cdot F(0.5 - \pi_i) > 0$, which is also therefore positive. (The two choices are conditionally independent because noise is drawn afresh each time.)

survey instability remains positive. Finally, although random utility can in a special case produce the same choice probabilities as probability matching,[8] they involve fundamentally different mechanisms for generating trial-to-trial variability within a respondent over time and hence testable predictions. All of these stochastic mechanisms are consistent with our claims about intrinsic human stochasticity.

Other Stochastic Decision Models Our claims do not depend on which decision making model is correct, so long as respondents are not making decisions solely deterministically. Indeed, decision models beyond random utility and probability matching also imply intrinsic stochasticity. These include models combining these stochastic approaches with the deterministic expected utility model, and others that introduce stochasticity directly (e.g., Stewart, Chater, and Brown, 2006; Graham, 2023). Similarly, preferences interpreted as "enduring personal dispositions" are observationally equivalent to those based on "evaluative judgments, formed when needed" (Schwarz, 2007a), including when formed as the average of "considerations" that happen to be salient at the moment of the survey (Zaller and Feldman, 1992), each of which may contribute to intrinsic stochasticity. All these and other decision making models imply intrinsic stochasticity that would lead to survey instability, as does our experiments which we now describe.

Empirical Evidence We now describe eight separate studies designed to hold respondents' preferences fixed and known to them by telling (and teaching) the respondent their beliefs (π_i) about their preferences. Although our claims only require the presence of some stochasticity in decision making, which would be satisfied under random utility, probability matching, or other stochastic models, we can provide some more information about the nature of the stochasticity. That is, if respondents are using the random utility model, η_{it} will be close to zero, their choices will follow the expected utility model, and survey instability should be close to zero. In contrast, if respondents are using probability matching or other more irreducibly stochastic models, then the results of our experiments

[8] When the random utility model is used for the logit or probit discrete choice statistical models, $F(\cdot)$ is assumed cumulative logistic or normal. If instead we assume a version of the linear probability discrete choice model, and so assume $\eta_{it} \sim \text{Uniform}(-0.5, 0.5)$, then because $F(u) = u + 0.5$, $\Pr(C_{it} = 1) = 1 - F(0.5 - \pi_i) = 1 - (0.5 - \pi_i + 0.5) = \pi_i$, which is probability matching. See also Vulkan (2000).

should yield high levels of stochasticity (and survey instability). That is, by teaching respondents their values of π_i, we greatly reduce respondents' *uncertainty* about the decision context that might otherwise generate stochasticity. It eliminates the possibility that instability is driven by *exact indifference*, which occurs when $\pi_i = 0.5$ under the expected utility model. And we also try to eliminate *approximate indifference*, which occurs when expected utility is violated due to a large enough error under the random utility model (e.g., $C_{it} = 1$ under expected utility when $\pi_i > 0.5$ but random utility can reverse when η_{it} is negative enough so that $\pi_i + \eta_{it} < 0.5$). (Alternatively, if η_{it} remains nonzero even after all identifiable sources of uncertainty are removed, then it cannot be interpreted as noise added to a deterministic process but is itself the irreducible stochasticity we describe.) As we now demonstrate, the results provide strong evidence that survey decision making is intrinsically stochastic (results which are consistent with additional analyses in Supplementary Appendix A4 which show that more divergent answer options produce less instability but nowhere near enough to explain the bulk of instability).

Figure 5 gives the results, with survey number on the horizontal axis and the percent choosing stochastically on the vertical axis. Point estimates range from 45% (Survey 31) to 59% (Survey 33) stochastic (as opposed to deterministic), with confidence intervals far from zero (which would have suggested people decide by rational choice or some other deterministic procedure).[9]

We begin with Survey 31 (at the left of Figure 5) which asks "For this election, say that you know that there is a three in four chance that Candidate A is better for you and the country, and a one in four chance that Candidate B is better. (Unfortunately, you don't know anything else about the candidates.)" We then ask the respondent to make four choices of A or B, for four separate offices in the same election. If respondents understand the question completely, the (stochastic) random utility model would reduce to the (deterministic) expected utility model, with no instability. Although mathematical perfection like this never translates to the empirical world, and respondents may even choose to limit their scarce attention to the question, our experiments are still designed to

[9]We also count as non-stochastic the smaller number who always do exactly the opposite of the rational choice model. Taken as a whole, the observable implications revealed in this figure are sample averages, not whether any one person is acting stochastically in any one situation.

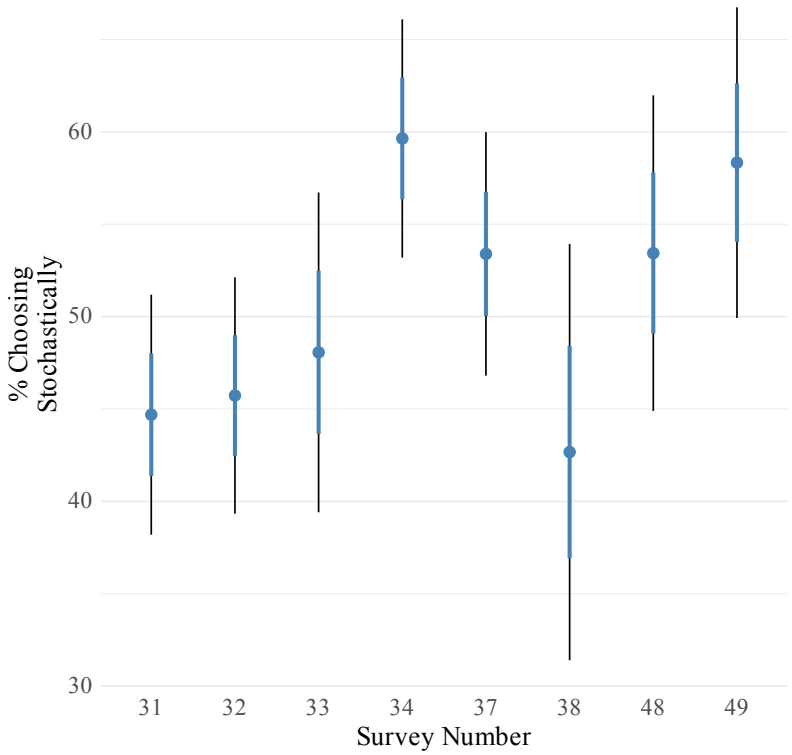


Figure 5: Percent Stochastic Decision Makers, across eight separate experiments.

reduce the variance of η_{it} to near 0. In this first experiment, the expected utility model predicts that all respondents would vote for Candidate A for all offices, but in fact an average of 45% of respondents violated the expected utility model, which is clear evidence of stochasticity.

In Surveys 32 and 33, we repeated the experiment, with the same setup, except that Candidate B was the preferred candidate (to rule out a deterministic decision rule by time optimizers). In Survey 34, we show that these results hold in a different context by asking respondents how they would vote in a single election if allowed to allocate ten votes freely between the candidates. This design assesses whether individuals add stochasticity to their decision making for a single time point (by splitting their votes between candidates against their maximum utility preferences). In Surveys 37 and 38, we repeated the experiment with three survey questions preceding the exercise where we tried to ensure that the respondents fully understood our instructions and gave them sufficient attention to do so, which we did by asking them about the instructions and informing them if they

gave the wrong answer. In Survey 49, we altered the probabilities of the candidate being better for the respondent and the country, to three in five and to nine in ten, respectively.

Across all eight surveys, an average of 50.5% of our sample violated deterministic mappings from beliefs to choices and revealed clear and robust evidence of enough stochasticity to constitute a meaningful component of observed survey instability.[10]

5.2 Cognitive Processes

The next most proximate factor generating intrinsic human stochasticity (from Figure 4), which then leads to survey instability, is cognitive processes. We now identify three observable implications of our framework in these processes that are particularly useful for social scientists and survey researchers, revealing how stochasticity produces instability in (1) survey questions with high cognitive complexity (complex or unfamiliarly worded questions and/or larger numbers of response categories), (2) respondents spending low time on the survey task for any given question, and (3) individuals with divergent cognitive processing for the two questions. Note that cognitive process (1) varies over questions, (2) over respondents within a question, and (3) over the two-trial pair of questions.

Cognitively Complex Survey Questions We show here that survey questions with higher cognitive complexity produce more instability, a relationship consistent with the human brain generating more stochasticity the more it is required to process a thought (Section 4). Cognitive complexity occurs, first, with more difficult-to-understand stem questions and, second, with more numerous response options. More response options typically require more mental effort to understand (and keep in working memory) and more thought to distinguish response categories from one another (see Tourangeau, 2021). Finally, questions on familiar topics asked of respondents with strongly held prior beliefs

[10]To convince oneself and others that probability matching is a common method of translating uncertain preferences into choices, and hence is an important source of survey instability, try presenting the following decision to an audience (composed of students, colleagues, or almost anyone; we have found it even works for an audience of academic decision scientists). Explain that everyone gets to vote in 10 elections, each for Candidate A or B. Unfortunately, all you know in each election is that Candidate A is better for you than Candidate B with 70% probability. By a show of hands, ask how many would vote for Candidate A in all 10 elections (usually just a few raise their hands). Then pick someone who didn't raise their hand and ask how they would vote and why, and in discussion you quickly hear something close to probability matching. We find this result robust and replicable, just like our formal experiments in this section.

requires less cognitive complexity.

We demonstrate this relationship by collecting a sample of widely used survey questions (from the Roper Center's iPoll service; see RoperCenter.cornell.edu/ipoll) spanning a large range of degrees of cognitive complexity and estimating the instability for each. To select questions, we used measures of cognitive complexity based on our own subjective evaluation, quantitative summaries (e.g., Lenzner, Kaczmirek, and Lenzner, 2010), and time-on-task (which a respondent spends reading, interpreting, and understanding the survey question) averaged over a sample, all with similar results.

In Figure 6, we plot the levels of survey instability (vertically) for each selected survey question by average time-on-task (horizontally). Results in blue show that instability increases steeply with higher levels of cognitive complexity. Our standard conjoint question is replicated in this analysis and highlighted near the middle of the pack (at about 15 seconds and 26% instability).[11]

The least cognitively complex question we included is gender (at the left): "What is your gender? (man, woman, other)". This is a simple question, with few response categories, on a highly familiar topic. As the figure indicates, this question has essentially zero instability, which is no surprise given that even small children can report their gender without delay and seemingly without any thought.

In contrast, consider the most cognitively complex question (Figure 6, top right):

> "Some people think we need much tougher government regulations on business in order to protect the environment. Suppose these people are at one end of a scale, at point 1. Others think that current regulations to protect the environment are already too much of a burden on business. Suppose these people are at the other end, at point 7. And, of course, some other people have opinions somewhere in between, at points 2, 3, 4, 5, or 6. Where would you place yourself on this scale, or haven't you thought much about this? (with choice buttons for 1–7)".

[11] To prevent our experimental protocol changes in this figure from violating the "no memory" condition, we included larger numbers of distractors than we ordinarily do with conjoints, and question types similar to the question of interest. We found that no respondent noticed and reported seeing a survey question twice. In addition, so that we can directly compare instability levels without risk of sample selection biases, we conducted this analysis by randomizing respondents to the eleven "treatment regimes" corresponding to the questions in Figure 6. See Supplementary Appendix A9 for details.

[12] We generated 1500 synthetic respondents using Claude by Anthropic to mirror our human respondents on demographic and political questions, location, race, age, education, income, party ID, ideology, and

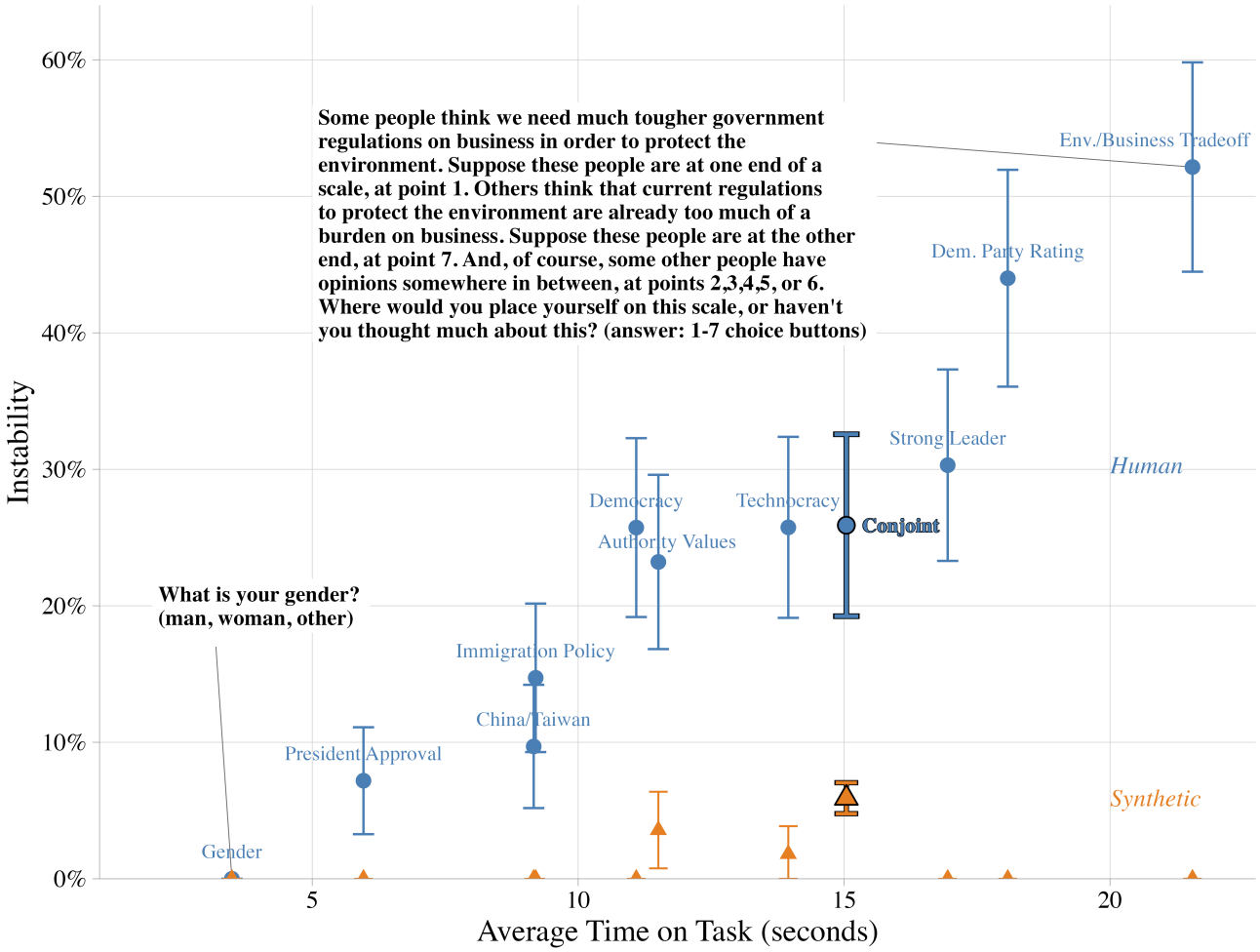


Figure 6: Survey Instability by Cognitive Complexity, measured by average time answering a survey question. Results (in blue) are averages for 11 survey questions with 95% confidence intervals from random samples of humans. Results at the bottom marked "synthetic" (in orange) are created by simulations from large language models, with well known problems of homogeneity, and confirm that our human sample is unlikely to be affected by bots.[12]

Think how much cognitive power it takes to understand this question. The respondent is being asked to visualize a sophisticated seven point scale that is clear at the start only in the researcher's mind (e.g., what is the precise semantic or conceptual difference between, say, points 4 and 5?). The scale is described carefully (although professorially), but requires sustained attention, considerable memory, and contextual understanding. It is about a topic that is unfamiliar to many. Question wording is also geared much more to-

2024 vote. We then administered the same survey to these simulated respondents as we did to the human respondents. As predicted from well-known flaws in large language models (Luo et al., 2026), instability levels were well below those from the human responses.

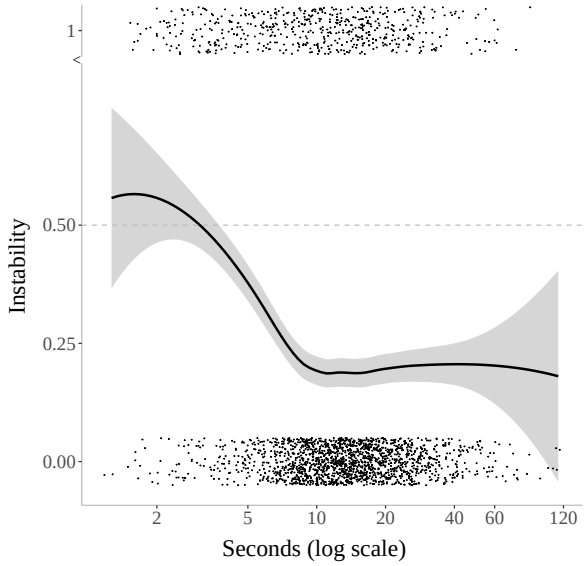
ward what the researcher wants than how many survey respondents think and talk in their day to day lives, making it even more complex for most respondents than researchers. The result is that survey instability on this question reaches a substantial 53%. See the Supplementary Appendix A9 for details on these and other questions.

The practical implication is clear: researchers can avoid some instability by writing survey questions respondents understand, in their language, the way they typically think about the world, rather than the way we researchers wish to analyze the resulting data. Lower the cognitive complexity of survey questions when possible, even if the resulting analysis may be complex, difficult, and sophisticated.

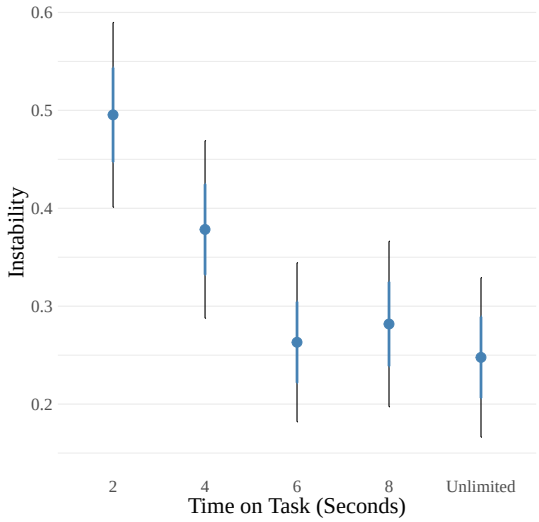
Low Time on Task In the previous section, we showed that across survey questions, *higher* average time-on-task measures higher cognitive complexity which yields higher instability. We now show that, for respondents within a survey question, *lower* time-on-task measures lack of attention to the question, which also generates higher instability. The distinction is stark but not contradictory, since understanding any question, no matter the level of cognitively complexity, requires some threshold degree of attention for understanding, and of course more for more complicated questions.

To study time on task for different respondents within questions, we conducted both observational and experimental studies, the former to understand reasonable parameter values for the experiment and the latter to demonstrate causality. We summarize our results in Figure 7, with the observational study in Panel (a) (Surveys 15–18, combined) and experimental study in (b) (Survey 20), and with measured time on task (averaged from Q1a and Q1b) on the horizontal axis and instability (D) on the vertical axis. Because we cannot know when respondents are paying attention or force them to pay attention, measured time on task in both cases is actually the maximum time on task that respondents could devote, with the true level possibly lower due to factors such as mind-wandering. Despite this unavoidable measurement error, a strong pattern can be seen in the figure, with instability for those who spend two seconds at about 0.5, the equivalent of flipping coins to choose a candidate, and dropping fast as respondents put more time in, until they have put in about 10 seconds, at which point instability levels off (at a still substantial

0.22) no matter how much more time the respondent puts in.



(a) Observational

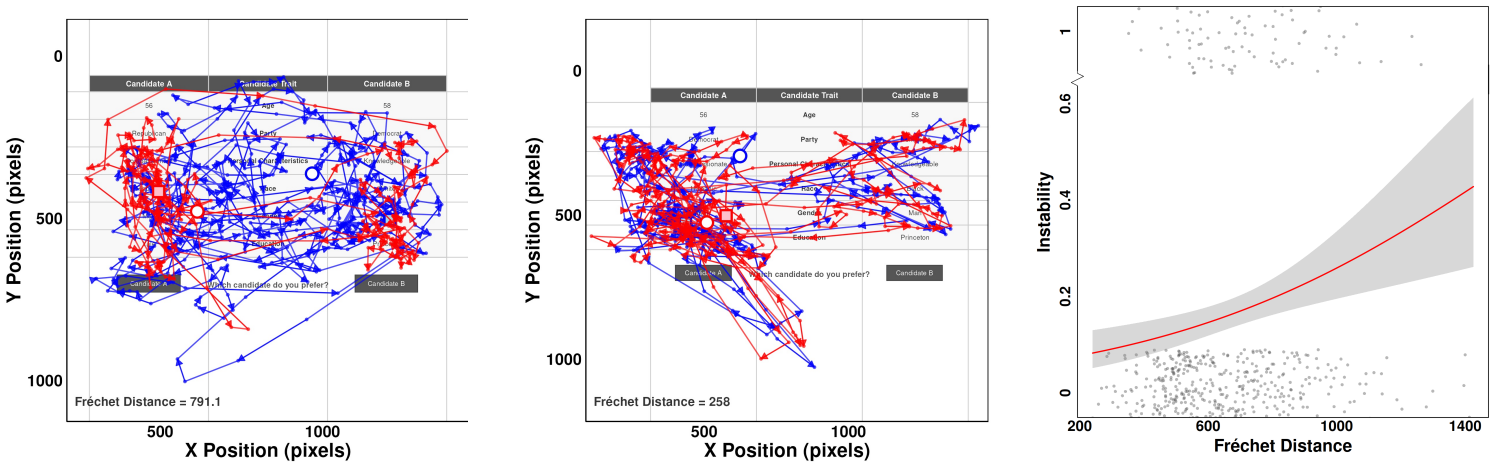


(b) Experimental

Figure 7: Time on Task Decreases Survey Instability, with observational results in (a) and a randomized experiment in (b). In (a), the dots are jittered indications for individual respondents of whether they gave the same (0) or different (1) answers for Q1a and Q1b, to which we have added a running mean (black line) and confidence interval (gray area around the line).

Our claim about the effects of time on task is causal, and so we take this a step further and run an experiment where respondents are randomly assigned time constraints of two, four, six, eight, or an unlimited number of seconds to complete each conjoint question (numbers we chose based on the results from Panel (a)). Again, these are maximum values of time on task. In Panel (b), we find results consistent with the observational study: Instability started at approximately 0.50 among those given only two seconds and dropped to 0.26 among respondents allotted six seconds.

Divergent Cognitive Processing In the previous two parts of this section, we show that survey instability can be generated by more cognitively complex survey questions and respondents who spend insufficient time reading and understanding the question. We now complement these results by demonstrating that, even when respondents spend sufficient time, instability can increase when processing information in the first and second questions in different ways. This "divergent cognitive processing," even when accompanied by



(a) Respondent Gaze Track (b) No Human Stochasticity (c) Distance by Survey Instability

Figure 8: Gaze Behavior via Eye Tracking (based on Surveys 55–56)

large amounts of time on task, can still increase instability. This finding may be the specific mechanism by which modern survey design, through appropriate priming reductions, can mitigate unnecessary instability.

We measure divergent cognitive processing via a well known source of intrinsic human stochasticity in the gaze behavior of human respondents (Vasudevan, Murthy, and Padhi, 2024). We do this via web-based eye tracking technology that we modified for these purposes and then demonstrate how this source of stochasticity helps explain survey instability. We do this in four steps.

First, to improve implementation accuracy for this set of experiments, we modify how we present the conjoint by listing attribute types in the middle, candidate A's attribute values to the left and B's to the right. Figure 8(a) gives an example of estimates of one respondent's gaze behavior for Q1a (in blue) and Q1b (in red), with starting points as circles, saccades as arrows, and fixations as dots. We print for reference the survey question in gray in the background.

The particular respondent portrayed in Figure 8(a) (who we chose for expository clarity) looked all over the survey form in Q1a (see the blue lines and dots) but, for Q1b, focused primarily on the attribute values for the candidates in Q1b (see the red lines and dots). The figure also vividly displays the stochasticity inherent in gaze behavior, which

can be seen (to a first order of approximation) by eye tracks darting around in apparently random local directions, even when forming global patterns.

Second, we summarize in a single number how different the gaze behavior is for Q1a and Q1b via the *Fréchet Distance* (Alt and Godau, 1995). This metric is most easily understood by first conceptualizing the two eye tracks as the paths of a person with a dog on a retractable leash. Each can move forward or stop any number of times, at any speed, following the existing path, but neither can move backwards. The Fréchet distance is the farthest the retractable leash has to extend at any point during their trip (for all possible trips following these rules). Because the gaze behavior is so different for Q1a and Q1b for the respondent in Figure 8(a), the Fréchet distance of 791 (displayed at the bottom left) is large, which can be seen because it is a substantial fraction of the pixels on either axis.[13]

Third, we take into account the fact that the variability observed in Figure 8(a) may be due to both the measurement device, which is important to remove, and intrinsic human stochasticity (the same point made in Section 3, although measurement error-induced stochasticity from eye tracking technology is likely larger than from survey instruments). To account for this issue, we construct what we call an artificial (nonstochastic) human being. We do this by videotaping a respondent answering a conjoint question and using a monitor playing the video in place of the real human to record data from the same eye tracking technology while "answering" Q1a and Q1b. That is, for Q1a and for Q1b, one computer monitor plays the video facing the other measuring the eye tracks. This enables us to eliminate all human stochasticity because the artificial person's eyes are doing exactly the same thing for both questions, a feat real (stochastic) humans cannot accomplish (regardless of how hard they try). In the absence of measurement error, the two curves would be identical (with a Fréchet distance of 0). We display results in Figure 8(b) where the red and blue eye tracings do track each other, but with noticeable variability. While the Fréchet distance for the one (extreme) actual person in Figure 8(a) is 791, the mean over a sample of ($n = 609$) respondents we collected is 653, and the one artificial respondent in Figure 8(b) is only 258 (repeats of the artificial respondent experiment gives

[13] Other common distance metrics for discrete paths, such as dynamic time warping or Hausdorff, give essentially the same substantive conclusions.

similar numerical results). We thus conclude that the mean artificial-respondent Fréchet distance of 258 places an upper bound on what eye-tracker noise alone can explain, so at least $(653-258)/653 \approx 60\%$ of the observed separation is attributable to intrinsic human variability, which is still quite substantial.

Finally, we arrive at the main point of this section, which is to show that divergent cognitive processing on the two questions, as measured by differences in a respondent's gaze behavior, predicts survey instability (Surveys 55-56). Figure 8(c) gives these results, with total Fréchet distance on the horizontal axis and survey instability on the vertical axis. Individual respondents appear as dots at 0 or 1, with jitter for graphic clarity. The relationship can be seen most clearly with the smoothed LOESS curve (in red, along with confidence intervals in gray), which indicates that survey instability is strongly predicted by the Fréchet distance. The true relationship is likely stronger than the graph indicates, assuming only that the measurement error-induced stochasticity we quantified above is random noise.

5.3 Psychological States

A third precursor in Figure 4 includes four different psychological states that generate stochasticity and influence survey instability — preoccupation, mind-wandering, persona, and attention. We present these precursors (and observable implications of our framework) in a typology in Table 1 organized by their common precursors — whether each is strongly influenced by other questions in the survey (in columns) or the respondent's conscious choice (in rows). In this section, we (a) describe these four psychological states along with how they are measured, (b) show how each influences survey instability (and time-on-task, which leads to instability), and (c) provide evidence that a measure of preoccupation used to filter out respondents unable to give a survey sufficient attention avoids the well known problems with attention check questions and meets essential criteria for good filters. Although preoccupation is our best practice recommendation, all these psychological states are highly related: For example, those who are preoccupied, mind-wandering, or take on certain persona give less attention to the survey. Understanding all four states is thus valuable, even though only preoccupation ultimately passes all criteria

for a good filter.

		Other Survey Questions: No	Other Survey Questions: Yes
Respondent's Conscious Choice	No	Preoccupation	Mind-wandering
	Yes	Persona	Attention

Table 1: A Typology of Psychological States that Affect Survey Instability, based on whether they are Strongly Influenced by Other Survey Questions (in columns) and Respondent's Conscious Choice (in rows).

Psychological States and Their Measures The most well-studied psychological state in survey research is *attention* (Table 1, bottom right), where respondents use their limited mental resources to focus on specific information (such as a survey question) while ignoring environmental stimuli or unrelated thoughts. The problem is that researchers must battle for respondents' attention, wielding only a survey with subjects of fascination to researchers. With attention strongly influenced by the content of the survey and the respondent's conscious choices (as emphasized by the table's construction), we often lose this battle, leading to higher levels of survey instability.

Attention is commonly measured by "attention check" questions that detect respondents that fall prey to simple tricks, such as instructions (added to ordinary questions) like "To show you are paying attention, ignore the following question and select 'Strongly disagree' no matter how you actually feel." or asking obvious questions such as "What color does a red pen write?". Unfortunately, researchers have shown that including these questions (to construct filters) can introduce rather than remove bias (Berinsky, Margolis, and Sances, 2014; Aronow, Baron, and Pinson, 2019; Varaine, 2023) in part because they change the survey situation from one of mutual cooperation, following the Gricean maxims of ordinary conversation (Schwarz, 1999), to one where respondents think they are being tricked and so treat other questions as a zero-sum game to quickly assess whether each question is an attention check and if not devote even less time than otherwise (Hauser and Schwarz, 2016). (Unfortunately, even explicit inducements provided by researchers, outside of financial payments, have had mixed results in increasing attention, increasing

accuracy, or improving response rates; Berinsky, Margolis, and Sances 2016; Singer and Ye 2013; Wolff 2019).

A second psychological state is *mind-wandering* (Table 1, top right), internally-focused thought that is predominantly self-referential and future-oriented (Baird, Smallwood, and Schooler, 2011) and which usually begins spontaneously via unconscious shifts in attention (Seli et al., 2016). Mind-wandering is endogenously affected by the survey but by definition not controlled by conscious thought (reflecting its position in the table). In the lab, respondents report mind-wandering about once every 80 seconds (Grandchamp, Braboszcz, and Delorme, 2014; Hasenkamp et al., 2012). Outside the lab, the average person spends *half* their waking day mind-wandering, with high levels of variation (Kane et al., 2017; Killingsworth and Gilbert, 2010). People report mind-wandering at least 30% of the time for every activity of daily life studied but one — sex, during which 10% admit to mind-wandering. Designing even more engaging surveys, with lower levels of mind-wandering, seems unlikely (and good luck trying to explain those research protocols to the IRB!).[14]

Unfortunately, the usual survey seems almost designed to encourage mind-wandering. Mind-wandering is more likely when people are bored, fatigued, and engaged in lengthy or repetitive tasks (Danckert and Merrifield, 2018; Walker and Trick, 2018) or who care more about their concerns and goals than a researcher's (Wong, Willoughby, and Machado, 2023). Consideration of discontinuous topics — as most surveys require — also increases mind-wandering (Smallwood, 2013). Instead of thinking of respondents as primarily interested in our survey questions and sometimes distracted, we should probably think of the survey as interrupting the mind-wandering respondents are doing as the survey commences, and so likely leads to survey instability through time-on-task limitations.[15]

[14]We also designed experiments to manipulate the prevalence of mind wandering but these all failed; see Supplementary Appendix A5. From our own introspection, we can suppress mind-wandering for extended periods only when serious bodily injury would result from even the briefest distraction, such as during a challenging ski run, but threatening the lives of research subjects would also not seem to be a good plan.

[15]During mind-wandering, the content of thought transitions more freely than in creative thinking or rumination (Christoff et al., 2016; Ciaramelli and Treves, 2019) and may also prime the respondent to answer a question in different ways. However, the topics respondents mind wander about usually differ enough from the research questions that fascinate social scientists. As a result, survey instability may arise more from time limitations than random priming (e.g., Girardeau et al., 2022). We confirmed this by asking a group of respondents who acknowledged mind-wandering for the content of their thoughts. We (and

We adapt the standard "probe-caught" methodology commonly used in psychology to measure spontaneous episodes of mind-wandering (Weinstein, 2018). To do this, we sometimes follow Q1a or Q1b with: "Everyone gets distracted at times. We are interested in how often people are distracted while taking surveys. While viewing the previous screen, did any thoughts come into your mind that were unrelated to the survey question, or did you only think about the survey question? Please be honest; it helps us understand!" Response options include (1) "Some thoughts unrelated to the survey question crossed my mind" and (2) "I only thought about the question posed there." Across eight probes in Surveys 2, 4, 13, 14, 16, 17, 18, and 26 (n = 2,138), 26.3% of respondents acknowledged mind-wandering some of the time. (We only ask about the previous screen and so this figure is a lower bound on total mind-wandering during the survey.)

The third psychological state is a *persona* (Table 1, bottom left), the self-identity people assign themselves when agreeing to participate in a survey, such as "earnest survey participants," "time optimizers," "whimsical randomizers," among others.

Although persona are well understood, they have not been the subject of extensive study in psychology and so we developed our own measure. Thus, we ask near the start of a survey: "How would you characterize yourself on the last survey you took? (1) I finished the survey as quickly as possible, (2) I balanced speed and giving thoughtful, accurate answers, (3) I answered thoughtfully and accurately, regardless of the time it took". By defining and measuring persona based on past behavior (and assuming some stable dispositional tendency), subsequent survey questions cannot have an effect on the measurement. However, the respondent can always self-consciously choose their own persona for this survey.

The final psychological state we consider is whether respondents are *preoccupied* before they start taking the survey (the top left box of Table 1). Those preoccupied are unusually focused or absorbed in thought about something other than the survey before

independently, a research assistant) coded each open ended response for whether the reported subject of their thoughts might conceivably prime or bias a respondent's answer. We were only able to find 3.5–5% of respondents' reported thoughts that might conceivably influence the answer. Moreover, some of these are irrelevant because, if a respondent had the same biasing thought during both questions, it would not affect instability. The main mechanism by which mind-wandering generates survey instability is thus likely through time-on-task limitations.

they begin, including topics such as social relationships, problems at work, financial difficulties, or health issues, etc. (Osborne and Gilbert, 1992). At times, all of us are preoccupied with something that leads to less ability to focus on anything else. Some groups are chronically preoccupied, such as those in poverty when thinking about money (Mani et al., 2013). Preoccupation is by definition (and by how we measure it) not affected by the content of the survey, difficult to control by the individual, and reduces cognitive capacity available for the survey. As a result, preoccupation typically reduces time on task, leading in turn to increased intrinsic stochasticity and thus survey instability.[16]

Through extensive informal experimentation and cognitive debriefing, we developed a simple survey question designed to measure preoccupation, with the goal of being widely understandable, personally meaningful, itself highly stable, not an inadvertent treatment, and useful for practical applications. The question follows:

> Everyone comes to a new task (like this survey) with things on their minds. How preoccupied were you, just before you started this survey, with thoughts or worries about your partner, job, health, children, friends, money, or other concerns? (1) Very preoccupied, (2) Fairly preoccupied, (3) Slightly preoccupied, or (4) Not at all preoccupied.

On average, across multiple surveys (Surveys 16, 26, 28, 52, and 53; $n = 926$), we find that a substantial 33.8% of respondents reported being very or fairly preoccupied.

Effect on Survey Instability and Time-on-Task We now offer Figure 9, which reveals the effect that all four psychological states have on survey instability and time-on-task. To explain the graphic analyses presented in parallel in the four panels, we begin with preoccupation (Panel a, top left). The horizontal axis of this graph indicates comparisons in the levels of preoccupation between "Not preoccupied at all" as the baseline and each of the other three higher levels as we move to the right. The top portion of the graph (in blue) shows the relationship with percent instability and the bottom with time-on-task, with respondents who are more preoccupied. The results are clear: Relative to

[16]The psychological literature addresses many related concepts, such as "cognitive busyness" (Osborne and Gilbert, 1992; Gilbert, Pelham, and Krull, 1988), "perceptual load," which affects visual processing, and "cognitive load," which affects how executive control functions are influenced by distractions (Lavie, 2010; Lavie et al., 2004) and the broader literature on "task-unrelated thoughts" (Savage, Potter, and Tatler, 2013; Forster and Lavie, 2009; Corbetta, Patel, and Shulman, 2008).

unpreoccupied respondents (with a baseline instability of 21% and 18.7 seconds on the survey task), those who reported being very preoccupied spent nearly five fewer seconds on the survey question and were a remarkable 13 percentage points more instable between Q1a and Q1b. The dose-response relationship is also clear in that the more preoccupied respondents are, the less time they spend on task and more instable they are.

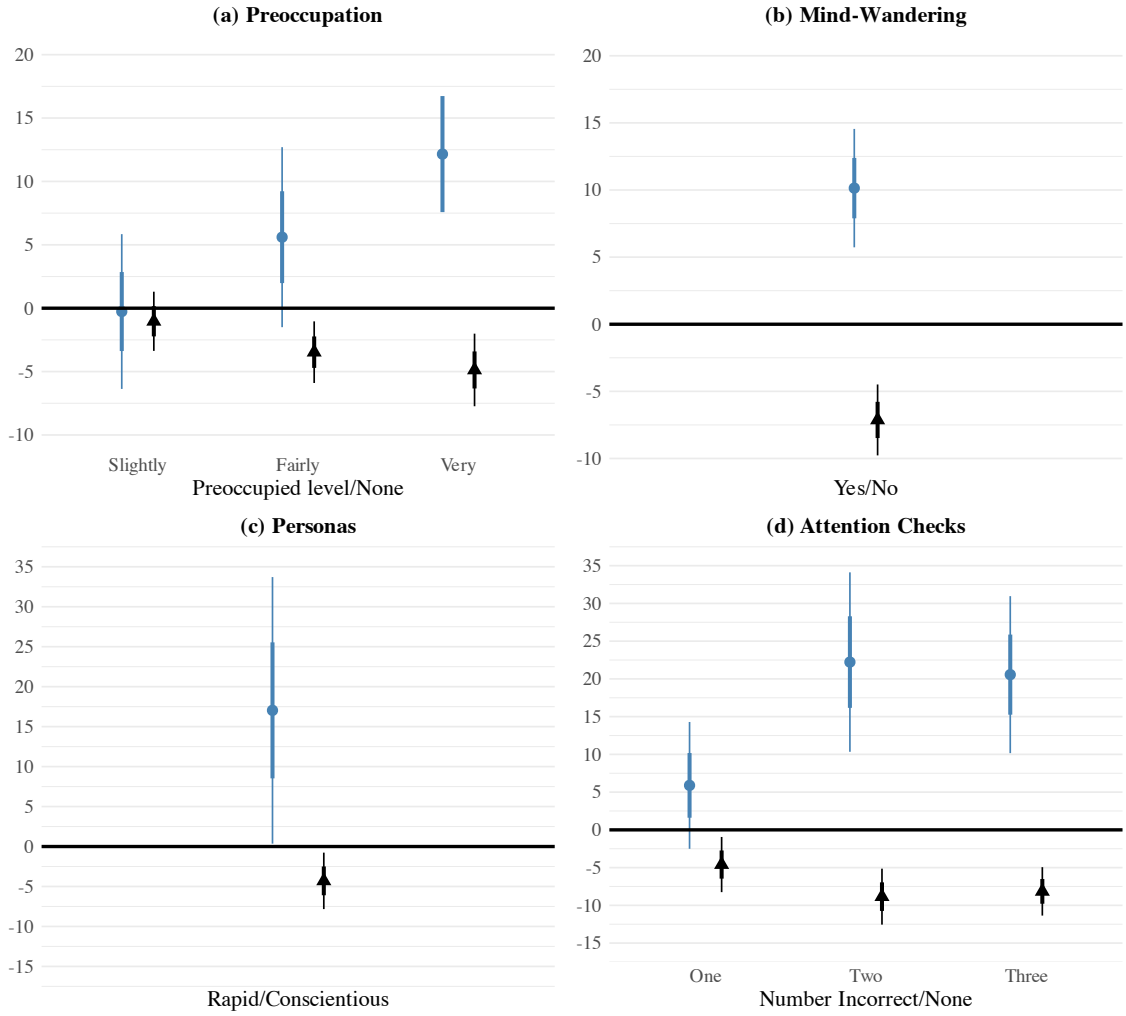


Figure 9: Effect of increasing levels of four psychological states (in four panels) on survey instability in percentage points (at the top of each, in blue) and time-on-task in seconds (at the bottom of each, in black). (See Surveys 16, 26, 28, 45, 52, and 53.)

The other three psychological states also increase survey instability and reduce time-on-task. Panel (b) shows that respondents who reported mind-wandering spent, on average, 6.75 fewer seconds actively thinking about the conjoint question compared to those who did not report mind-wandering (i.e., relative to a 15.5 second baseline). In addition,

these individuals exhibited 10.1 percentage points more survey instability (relative to a 21% baseline). Panel (c) reveals that changing from the conscientious persona (response category 2 or 3) to a rapid responder persona (category 1) is associated with large increases in survey instability (in blue) and reduction in time on task (in black). Finally, Panel (d) shows the ability of attention check questions to predict survey instability, which has not been studied directly before. The result shows that failing attention check questions on average does predict instability (in blue) and that may be due, in part, to the reductions in time-on-task (in black).

Preoccupation as an Improved Filter Filtering respondents may be necessary but it is a risky step that, when wrong, can induce selection bias (by removing observations related to variables of interest) or measurement-induced bias (where questions added to build filters bias responses to subsequent questions, special cases of which include priming, question order effects, reactivity, context effects, and, in experiments, post-treatment bias). Selection bias is best addressed in this context by how we analyze the filtered data (such as by changing the quantity of interest, imputing, or oversampling the filtered group; see Section 6). In contrast, measurement-induced bias can only feasibly be addressed by choosing better filter questions. Good filter questions (1) are not influenced by other questions in the survey; (2) are not influenced by the respondent's conscious choices; (3) distinguish between respondents with different levels of survey instability; (4) do not influence other questions in the survey (by inadvertently becoming a treatment variable), and (5) are more stable over time than the variables of interest they seek to explain.

By the construction of Table 1, only preoccupation (in the top-left cell) satisfies criteria (1) and (2). Figure 9(a) shows that preoccupation also satisfies (3). In the rest of this section, we show how preoccupation also satisfies (4) and (5).

To evaluate (4), we estimate causal effects by running randomized experiments with and without the preoccupation question (and, separately, the three others) and evaluating its effect. Figure 10(a) shows that preoccupation (at the top), as well as the other variables, do not have measurable treatment effects on levels of instability.[17]

[17]Future researchers could also test whether including preoccupation effects the substantive variable of

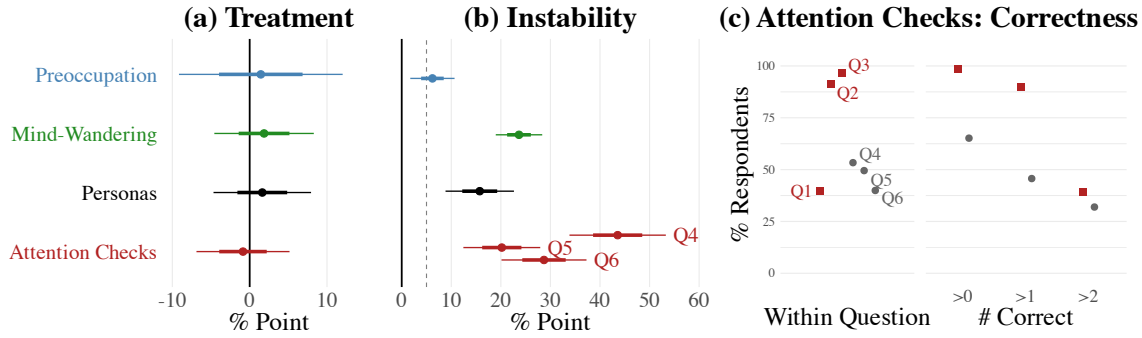


Figure 10: Filter Evaluations. Panel (a) shows the lack of treatment effect for all four psychological states; Panel (b) reports the instability of filters based on each psychological state, and Panel (c) shows the huge variability in different attention check questions. (See Surveys 16, 26, 28, 45, 52, and 53.)

We assess (5) by asking each of the filters at the start and end of the survey (e.g., to measure the instability of preoccupation we used this design in Survey 52: S P1 Q1a D1 D2 D3 Q1b P2). Figure 10(b) summarizes the results. Although average instability in typical questions is about 25%, mind-wandering, persona, and three versions of attention check questions (in green, black, and red, respectively) are not measurably lower and so would not make useful filters. In contrast, preoccupation (in blue at the top of (b)) is highly stable, with only 6.2% of respondents changing their answers.

Finally, we designed Figure 10(c) to study the levels of variation across filters created based on different commonly used attention check questions. We conduct two surveys (Survey 45 in red and 53 in gray) using six different individual attention check questions, and report (at the left) the percent passing the check, which range from about 30% to nearly 100%. Because some scholars include multiple attention check questions, we also show the variability given 1, 2, or 3 correct, which has a slightly larger range. Because of this extremely wide variability, it would not be clear what to use (even if, contrary to Panel (b), we had found they were stable over the survey).[18]

These results for survey instability reinforce the conclusion in the literature of others

interest in particular studies. We also note that the lack of a treatment effect in our sample for attention checks (Panel (a), in red) is different from prior research, possibly because their high prevalence in so many surveys has led even control group respondents to expect them (which would be worrisome for the future of survey research).

[18]We might think that we could choose a passing threshold ex post, say 20% of respondents with the worst scores could be filtered out. However, with varying but unobserved true levels of instability across surveys and subgroups, this procedure would not accomplish much of value.

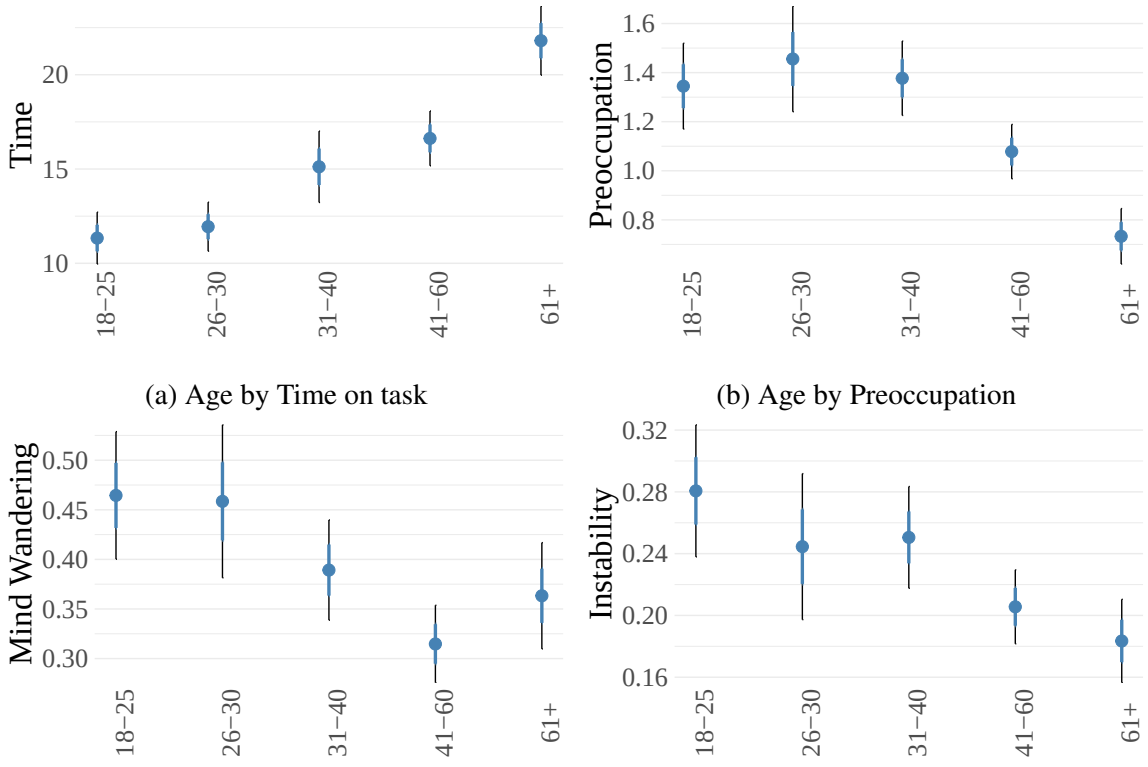
who have studied this issue: most attention check questions should not be used to filter out respondents. They also support the idea that preoccupation is, at least over the course of a survey, a highly stable characteristic reflecting an individual's absorption in particular thoughts or concerns. The preoccupation item is unlikely to influence respondents' answers to other questions, but it does influence instability; adding this single question to a survey has low cost without much downside. This of course makes sense: No matter how fascinating a survey may be to us researchers, it is no match for the attention of a respondent having a fight with their spouse or dealing with a major financial problem.

5.4 Individual Characteristics

The most distal contributor to survey instability we consider in Figure 4 is individual characteristics, the relatively fixed features of people such as demographics, socioeconomics, or prior political knowledge, that generate different levels of stochasticity relative to their signal. For example, Freeder, Lenz, and Turney (2019) finds that voters who understand issues in the same way as their political party have lower survey instability. Understanding which individual characteristics drive instability is of direct substantive interest. It is also of importance methodologically because some subgroups may be much more difficult to study, subject to larger biases in trying to understand them, and simultaneously of potential use as filters because they are all exogenous, measurable prior to the survey at hand, and not a function of the respondent's choices during the survey.

As just one example of individual characteristics that may lead to survey instability, we look at the effects of age and report some surprising results. In Figure 11, we examine time-on-task (TOT), preoccupation, mind-wandering, and survey instability (vertical axis of the four panels) by age group (horizontal axis of all four).

As is well known, as people age, cognitive capacities can decline. The results in Figure 11a, indicating that older respondents take more time responding to survey questions, seem consistent with this finding. However, as Figure 11b shows, older survey respondents tend to be far less preoccupied than the young. Older respondents possibly understand their weaknesses and more than compensate by mind-wandering far less (Figure 11c; see also Jordão et al. 2019). In total, perhaps as a result of these patterns, we



(a) Age by Time on task

(b) Age by Preoccupation

(c) Age by Mind-wandering

(d) Age by Survey Instability

Figure 11: Age Predicts Time on task, Preoccupation, Mind-wandering, and Instability

also see survey instability dropping by a considerable 40% (11 percentage points) from the youngest to the oldest age groups (Figure 11d).

Age of course is just one illustration of what we might learn by studying how survey instability varies with individual characteristics. This result tells us a great deal about our respondents and issues studying them, but it of course cannot be used to filter out respondents without introducing various biases, a subject to which we now turn.

6 Substantive Conclusions, Practical Implications

Human beings have clear preferences on many issues of the day, rather than forming them whimsically, and survey researchers do not usually use massively flawed survey instruments. Instead, a large portion of survey instability appears to be due to intrinsic human stochasticity, a fundamental feature of human beings that has been largely undiscussed in prior survey research. This stochasticity can sometimes be influenced but cannot be

removed. Intrinsic human stochasticity, like prior explanations for survey instability, is a conceptual theory, and so our evidence comes from data we created about observable implications of the theory. Numerous other observable implications can also be tested, and we hope scholars will follow up on the trails we may have left. We hope our results, and the sources of instability at various levels of proximity we identified, will enable scholars to better understand the populations they are studying, and that the following resulting proposed adjustments to survey design and analysis practices can be used to improve the inferences we all make from survey data.

We now conclude with three areas our results suggest for adjustments in "best practices" for survey research to deal with survey instability. First, to *improve survey instruments*, our findings suggest designing the survey to reduce the cognitive complexity of questions and responses. Regardless of how sophisticated our theories and concepts may be, respondents are not scholars and should not be expected to see the world the way we do or to understand our research concepts. We need to meet respondents where they are and ask information available to them. An easy, unobtrusive measure of cognitive complexity worth using is the time respondents take on average to answer a survey question (such as in a small pretest survey). We can also improve survey instruments by reducing priming effects that may lead to divergent cognitive processing, bias, and survey instability — something that should become easier as eye tracking technology becomes easier to use.

We also strongly recommend that researchers measure their own survey's instability, D, whenever possible, by asking their key substantive outcome variable a second time in the same survey. This provides important additional information about our research subjects. Expanding our quantities of interest from the mean to the variance also opens up novel research questions, such as understanding how the stability of different subgroups in different issue areas. In some situations, researchers may wish to use D to estimate choice fidelity, $\Pr(C_{it} = \rho_i)$, and use it under a "swapping error" interpretation of probability matching to correct (Clayton et al., 2025).

Second, to *identify respondents to filter out*, we recommend measuring both preoccu-

pation and time-on-task. Our single-question measure of preoccupation strongly predicts survey instability and has excellent statistical properties (it is highly stable over the course of the survey, does not disrupt answers to other questions, and does not seem possible to even intentionally change by either the respondent or researcher). Time on task involves no additional questions, but two implementation questions are relevant. In one, because measuring true TOT directly is difficult, we suggest measuring total TOT, and using it to determine a value below which respondents could not have given serious consideration to the question (six seconds worked for our conjoint questions for example). Those above the threshold may include some that were mind-wandering or time optimizing but those below it almost surely should be filtered out. In the other, to avoid endogeneity bias, time-on-task should be measured for a few "burn in" questions prior to those of interest rather than on the question of interest. We find time on task to be highly correlated across questions, and so this exogenous measure typically both avoids bias and can be helpful in identifying those not giving the survey question the attention needed.

Finally, to *conduct proper statistical analyses*, if we filter out some respondents, and take no other action then we must openly acknowledge that our quantities of interest have changed and now only apply to certain subgroups. The groups that are preoccupied or spend less than adequate time on task should be explicitly identified (with analyses such as in Figure 11b). For questions where we are unwilling to change the quantity of interest (say for trying to predict the winner in an upcoming election), researchers must find some other source of data on the preoccupied group. This information could potentially be imputed from the same survey, as in studies of missing data, or survey researchers could oversample (or sample in different ways) the demographic groups that are most likely to be preoccupied.[19] A final possibility is to keep those preoccupied in the analysis but to analyze them separately, and with considerable caution.

Even if we are able to avoid or work around some instability by better survey design and more by appropriately filtering out respondents who would have unacceptably high levels, the intrinsic human stochasticity of the creatures we study will still have instability

[19] An alternative approach might be to use "principal stratification" which involves finding exogenous predictors of endogenous variables, making specific modeling assumptions, and then avoiding bias by conditioning on functions of these variables (VanderWeele, 2011).

levels that should not be ignored. The numerous remaining aspects of instability should be studied as an essential feature of our subjects of interest.

References

Achen, Christopher H (1975). "Mass political attitudes and the survey response". In: *American Political Science Review* 69.4, pp. 1218–1231.

Alt, Helmut and Michael Godau (1995). "Computing the Fréchet distance between two polygonal curves". In: *International Journal of Computational Geometry & Applications* 5.01n02, pp. 75–91.

Alvarez, R Michael, Lonna Rae Atkeson, Ines Levin, and Yimeng Li (2019). "Paying attention to inattentive survey respondents". In: *Political Analysis* 27.2, pp. 145–162.

Alwin, Duane F and Jon A Krosnick (1991). "The reliability of survey attitude measurement: The influence of question and respondent attributes". In: *Sociological methods & research* 20.1, pp. 139–181.

Ansolabehere, Stephen, Jonathan Rodden, and James M. Snyder (2008). "The Strength of Issues: Using Multiple Measures to Gauge Preference Stability, Ideological Constraint, and Issue Voting". In: *American Political Science Review* 102.02, pp. 215–232. DOI: 10.1017/S0003055408080210.

Arias, Victor B, Fernando P Ponce, Luis E Garrido, Maria Dolores Nieto-Canaveras, Agustin Martinez-Molina, and Benito Arias (2023). "Detecting non-content-based response styles in survey data: An application of mixture factor analysis". In: *Behavior Research Methods*, pp. 1–17.

Aronow, Peter M, Jonathon Baron, and Lauren Pinson (2019). "A note on dropping experimental subjects who fail a manipulation check". In: *Political Analysis* 27.4, pp. 572–589.

Backström, Kim, Alexandru Cernat, Rasmus Sirén, and Peter Söderlund (2025). "Measurement error when surveying issue positions: a MultiTrait MultiError approach". In: *Political Science Research and Methods*, pp. 1–18.

Baird, Benjamin, Jonathan Smallwood, and Jonathan W Schooler (2011). "Back to the future: Autobiographical planning and the functionality of mind-wandering". In: *Consciousness and cognition* 20.4, pp. 1604–1611.

Berinsky, Adam J (2017). "Measuring public opinion with surveys". In: *Annual review of political science* 20, pp. 309–329.

Berinsky, Adam J, Michele F Margolis, and Michael W Sances (2014). "Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys". In: *American journal of political science* 58.3, pp. 739–753.

— (2016). "Can we turn shirkers into workers?" In: *Journal of Experimental Social Psychology* 66, pp. 20–28.

Blackwell, Matthew, Jacob R Brown, Sophie Hill, Kosuke Imai, and Teppei Yamamoto (2025). "Priming bias versus post-treatment bias in experimental designs". In: *Political Analysis* 33.4, pp. 361–377.

Brady, Henry E. (Apr. 1993). "Why Wiley-Wiley Models Do Not Tell You Enough". In: Midwest Political Science Association, Chicago.

Brooks, Jonathan CW, Olivia K Faull, Kyle TS Pattinson, and Mark Jenkinson (2013). "Physiological noise in brainstem FMRI". In: *Frontiers in human neuroscience* 7, p. 623.

Bryant, Penelope A, Gordon K Smyth, Roy Robins-Browne, and Nigel Curtis (2011). "Technical variability is greater than biological variability in a microarray experiment but both are outweighed by changes induced by stimulation". In: *PloS one* 6.5, e19556.

Caron, Sophie JC, Vanessa Ruta, Larry F Abbott, and Richard Axel (2013). "Random convergence of olfactory inputs in the Drosophila mushroom body". In: *Nature* 497.7447, pp. 113–117.

Christoff, Kalina, Zachary C Irving, Kieran CR Fox, R Nathan Spreng, and Jessica R Andrews-Hanna (2016). "Mind-wandering as spontaneous thought: a dynamic framework". In: *Nature reviews neuroscience* 17.11, pp. 718–731.

Ciaramelli, Elisa and Alessandro Treves (2019). "A mind free to wander: neural and computational constraints on spontaneous thought". In: *Frontiers in psychology* 10, p. 39.

Clayton, Katherine, Yusaku Horiuchi, Aaron R Kaufman, Gary King, and Mayya Komisarchik (2025). "Correcting Measurement Error Bias in Conjoint Survey Experiments". In: *American Journal of Political Science [forthcoming]*. URL: GaryKing.org/conjointE.

Converse, Philip E (1964). "The nature of belief systems in mass publics (republished 2006)". In: *Critical review* 18.1-3, pp. 1–74.

Corbetta, Maurizio, Gaurav Patel, and Gordon L Shulman (2008). "The reorienting system of the human brain: from environment to theory of mind". In: *Neuron* 58.3, pp. 306–324.

Danckert, James and Colleen Merrifield (2018). "Boredom, sustained attention and the default mode network". In: *Experimental brain research* 236, pp. 2507–2518.

Erikson, Robert S (1979). "The SRC panel data and mass political attitudes". In: *British Journal of Political Science* 9.1, pp. 89–114.

Forster, Sophie and Nilli Lavie (2009). "Harnessing the wandering mind: The role of perceptual load". In: *Cognition* 111.3, pp. 345–355.

Freeder, Sean, Gabriel S Lenz, and Shad Turney (2019). "The importance of knowing 'what goes with what': Reinterpreting the evidence on policy attitude stability". In: *The Journal of Politics* 81.1, pp. 274–290.

Gilbert, Daniel T, Brett W Pelham, and Douglas S Krull (1988). "On cognitive busyness: When person perceivers meet persons perceived." In: *Journal of personality and social psychology* 54.5, p. 733.

Gilbert, Daniel T, Elizabeth C Pinel, Timothy D Wilson, Stephen J Blumberg, and Thalia P Wheatley (1998). "Immune neglect: a source of durability bias in affective forecasting." In: *Journal of personality and social psychology* 75.3, p. 617.

Girardeau, Jean-Charles, Marco Sperduti, Philippe Blondé, and Pascale Piolino (2022). "Where is my mind...? The link between mind wandering and prospective memory". In: *Brain sciences* 12.9, p. 1139.

Godefroid, Marie-E, Ralf Plattfaut, and Björn Niehaves (2023). "How to measure the status quo bias? A review of current literature". In: *Management Review Quarterly* 73.4, pp. 1667–1711.

Graham, Matthew H (2023). "Measuring misperceptions?" In: *American Political Science Review* 117.1, pp. 80–102.

Grandchamp, Romain, Claire Braboszcz, and Arnaud Delorme (2014). "Oculometric variations during mind wandering". In: *Frontiers in psychology* 5, p. 31.

Groves, Robert M and Lars Lyberg (2010). "Total survey error: Past, present, and future". In: *Public opinion quarterly* 74.5, pp. 849–879.

Hajcak, Greg, Alexandria Meyer, and Roman Kotov (2017). "Psychometrics and the neuroscience of individual differences: Internal consistency limits between-subjects effects." In: *Journal of Abnormal Psychology* 126.6, p. 823.

Hasenkamp, Wendy, Christine D Wilson-Mendenhall, Erica Duncan, and Lawrence W Barsalou (2012). "Mind wandering and attention during focused meditation: a fine-grained temporal analysis of fluctuating cognitive states". In: *Neuroimage* 59.1, pp. 750–760.

Hauser, David J and Norbert Schwarz (2016). "Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants". In: *Behavior research methods* 48.1, pp. 400–407.

Heams, Thomas (2014). "Randomness in biology". In: *Mathematical Structures in Computer Science* 24.3, e240308.

Holland, Paul W. (1986). "Statistics and Causal Inference". In: *Journal of the American Statistical Association* 81, pp. 945–960.

Hopkins, Daniel and Gary King (2010). "Improving Anchoring Vignettes: Designing Surveys to Correct Interpersonal Incomparability". In: *Public Opinion Quarterly*, pp. 1–22. URL: http://j.mp/jVFIVg.

Hout, Michael and Orestes P Hastings (2016). "Reliability of the core items in the General Social Survey: estimates from the three-wave panels, 2006–2014". In: *Sociological Science* 3, pp. 971–1002.

Huang, Jason L, Paul G Curran, Jessica Keeney, Elizabeth M Poposki, and Richard P DeShon (2012). "Detecting and deterring insufficient effort responding to surveys". In: *Journal of Business and Psychology* 27, pp. 99–114.

Ilan, Yaron (2020). "Order through disorder: the characteristic variability of systems". In: *Frontiers in Cell and Developmental Biology* 8, p. 495391.

Isenberg, Naomi and Markus Brauer (2022). "Commitment and Consistency". In: *The Routledge Research Encyclopedia of Psychology Applied to Everyday Life*. Ed. by R.A.R. Gurung. Routledge.

Jenke, Libby, Kirk Bansak, Jens Hainmueller, and Dominik Hangartner (2021). "Using Eye-Tracking to Understand Decision-Making in Conjoint Experiments". In: *Political Analysis* 29.1, pp. 75–101.

Jordão, Magda, Fernando Ferreira-Santos, Maria Salomé Pinho, and Peggy L St Jacques (2019). "Meta-analysis of aging effects in mind wandering: Methodological and sociodemographic factors." In: *Psychology and aging* 34.4, p. 531.

Kane, Michael J, Georgina M Gross, Charlotte A Chun, Bridget A Smeekens, Matt E Meier, Paul J Silvia, and Thomas R Kwapil (2017). "For whom the mind wanders, and when, varies across laboratory and daily-life settings". In: *Psychological science* 28.9, pp. 1271–1289.

Kang, Myong-Il and Shinsuke Ikeda (2014). "Time discounting and smoking behavior: evidence from a panel survey". In: *Health economics* 23.12, pp. 1443–1464.

Keller, L Robin (1992). "Properties of utility theories and related empirical phenomena". In: *Utility theories: Measurements and applications*, pp. 3–23.

Killingsworth, Matthew A. and Daniel T. Gilbert (2010). "A wandering mind is an unhappy mind". In: *Science* 330.6006, pp. 932–932.

King, Ella M, Megan C Engel, Caroline Martin, Alp M Sunol, Qian-Ze Zhu, Sam S Schoenholz, Vinothan N Manoharan, and Michael P Brenner (2025). "Inferring interaction potentials from stochastic particle trajectories". In: *Physicial Review Research (arXiv preprint arXiv:2406.01522)*.

King, Gary (1989). *Unifying Political Methodology: The Likelihood Theory of Statistical Inference*. Ann Arbor: Michigan University Press.

— (Sept. 1995). "Replication, Replication". In: *PS: Political Science and Politics* 28.3, pp. 443–499. URL: j.mp/jCyfF1.

— (2007). "An Introduction to the Dataverse Network as an Infrastructure for Data Sharing". In: *Sociological Methods and Research* 36.2. http://gking.harvard.edu/files/abs/dvn-abs.shtml, pp. 173–199.

King, Gary, James Honaker, Anne Joseph, and Kenneth Scheve (Mar. 2001). "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation". In: *American Political Science Review* 95.1. http://j.mp/lsZDuW, pp. 49–69.

King, Gary, Robert O. Keohane, and Sidney Verba (1994). *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton: Princeton University Press. URL: bit.ly/gKroKsV.

King, Gary, Christopher J.L. Murray, Joshua A. Salomon, and Ajay Tandon (Feb. 2004). "Enhancing the Validity and Cross-cultural Comparability of Measurement in Survey Research". In: *American Political Science Review* 98.1. GaryKing.org/files/abs/vign-abs.shtml, pp. 191–207.

Kustov, Alexander, Dillon Laaker, and Cassidy Reller (2021). "The stability of immigration attitudes: Evidence and implications". In: *The Journal of Politics* 83.4, pp. 1478–1494.

Lavie, Nilli (2010). "Attention, distraction, and cognitive control under load". In: *Current directions in psychological science* 19.3, pp. 143–148.

Lavie, Nilli, Aleksandra Hirst, Jan W De Fockert, and Essi Viding (2004). "Load theory of selective attention and cognitive control." In: *Journal of experimental psychology: General* 133.3, p. 339.

Lazarsfeld, Paul F (1948). "The use of panels in social research". In: *Proceedings of the American Philosophical Society* 92.5, pp. 405–410.

Lazarsfeld, Paul F., Bernard R. Berelson, and William N. McPhee (1948). *Elmira Community Study*. ICPSR [distributor] [2006]. DOI: 10.3886/ICPSR07200.v1.

Lee, Leonard, On Amir, and Dan Ariely (2009). "In search of homo economicus: Cognitive noise and the role of emotion in preference consistency". In: *Journal of consumer research* 36.2, pp. 173–187.

Lee, Leonard, Michelle P Lee, Marco Bertini, Gal Zauberman, and Dan Ariely (2015). "Money, time, and the stability of consumer preferences". In: *Journal of Marketing Research* 52.2, pp. 184–199.

Lenzner, Timo, Lars Kaczmirek, and Alwine Lenzner (2010). "Cognitive burden of survey questions and response times: A psycholinguistic experiment". In: *Applied cognitive psychology* 24.7, pp. 1003–1020.

Liang, Guanxiang and Frederic D Bushman (2021). "The human virome: assembly, composition and host interactions". In: *Nature Reviews Microbiology* 19.8, pp. 514–527.

Lim, Chaeyoon, Carol Ann MacGregor, and Robert D Putnam (2010). "Secular and liminal: Discovering heterogeneity among religious nones". In: *Journal for the Scientific Study of Religion* 49.4, pp. 596–618.

Lo, Andrew W, Katherine P Marlowe, and Ruixun Zhang (2021). "To maximize or randomize? An experimental study of probability matching in financial decision making". In: *Plos one* 16.8, e0252540.

Luo, Queenie, Gary King, Michael Puett, and Michael D. Smith (2026). "Inducing Sustained Creativity and Diversity in Large Language Models". In.

Mani, Anandi, Sendhil Mullainathan, Eldar Shafir, and Jiaying Zhao (2013). "Poverty impedes cognitive function". In: *science* 341.6149, pp. 976–980.

McDonnell, Mark D, Joshua H Goldwyn, and Benjamin Lindner (2016). *Neuronal stochastic variability: influences on spiking dynamics and network activity*.

McDonnell, Mark D and Lawrence M Ward (2011). "The benefits of noise in neural systems: bridging theory and experiment". In: *Nature Reviews Neuroscience* 12.7, pp. 415–425.

McGrath, Robert E, Matthew Mitchell, Brian H Kim, and Leaetta Hough (2010). "Evidence for response bias as a source of error variance in applied assessment." In: *Psychological bulletin* 136.3, p. 450.

Meade, Adam W and S Bartholomew Craig (2012). "Identifying careless responses in survey data." In: *Psychological methods* 17.3, p. 437.

Mercier, Hugo (2020). *Not born yesterday: The science of who we trust and what we believe*. Princeton University Press.

— (2022). "Confirmation bias–myside bias". In: *Cognitive illusions*. Routledge, pp. 78–91.

Molloy, Mark P, Erin E Brzezinski, Junqi Hang, Michael T McDowell, and Ruth A VanBogelen (2003). "Overcoming technical variation and biological variation in quantitative proteomics". In: *Proteomics* 3.10, pp. 1912–1919.

Osborne, Randall E and Daniel T Gilbert (1992). "The preoccupational hazards of social life." In: *Journal of Personality and Social Psychology* 62.2, p. 219.

Pinsoneault, Terry B (2007). "Detecting random, partially random, and nonrandom Minnesota Multiphasic Personality Inventory-2 protocols." In: *Psychological assessment* 19.1, p. 159.

Preuss, Malte (2021). "Intra-individual stability of two survey measures on forward-looking attitude". In: *Journal of Economic Behavior & Organization* 190, pp. 201–227.

Rentfrow, Peter J, Lewis R Goldberg, and Daniel J Levitin (2011). "The structure of musical preferences: a five-factor model." In: *Journal of personality and social psychology* 100.6, p. 1139.

Rolls, Edmund T and Gustavo Deco (2010). *The noisy brain: stochastic dynamics as a principle of brain function*. Oxford university press.

Rose, Katie A, Mehdi Molaei, Michael J Boyle, Daeyeon Lee, John C Crocker, and Russell J Composto (2020). "Particle tracking of nanoparticles in soft matter". In: *Journal of Applied Physics* 127.19.

Samuelson, William and Richard Zeckhauser (1988). "Status quo bias in decision making". In: *Journal of risk and uncertainty* 1, pp. 7–59.

Savage, Steven W, Douglas D Potter, and Benjamin W Tatler (2013). "Does preoccupation impair hazard perception? A simultaneous EEG and eye tracking study". In: *Transportation research part F: traffic psychology and behaviour* 17, pp. 52–62.

Schaeffer, Nora Cate and Jennifer Dykema (2020). "Advances in the science of asking questions". In: *Annual Review of Sociology* 46, pp. 37–60.

Schwarz, Norbert (1999). "Self-Reports: How the Questions Shape the Answers". In: *American Psychologist* 54.2, pp. 93–105.

— (2007a). "Attitude construction: Evaluation in context". In: *Social cognition* 25.5, pp. 638–656.

— (2007b). "Cognitive aspects of survey methodology". In: *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition* 21.2, pp. 277–287.

Seli, Paul, Evan F Risko, Daniel Smilek, and Daniel L Schacter (2016). "Mind-wandering with and without intention". In: *Trends in cognitive sciences* 20.8, pp. 605–617.

Sender, Ron, Shai Fuchs, and Ron Milo (2016). "Revised estimates for the number of human and bacteria cells in the body". In: *PLoS biology* 14.8, e1002533.

Sender, Ron and Ron Milo (2021). "The distribution of cellular turnover in the human body". In: *Nature medicine* 27.1, pp. 45–48.

Singer, Eleanor and Cong Ye (2013). "The use and effects of incentives in surveys". In: *The ANNALS of the American Academy of Political and Social Science* 645.1, pp. 112–141.

Smallwood, Jonathan (2013). "Distinguishing how from why the mind wanders: a process–occurrence framework for self-generated mental activity." In: *Psychological bulletin* 139.3, p. 519.

Sosulski, Dara L, Maria Lissitsyna Bloom, Tyler Cutforth, Richard Axel, and Sandeep Robert Datta (2011). "Distinct representations of olfactory information in different cortical centres". In: *Nature* 472.7342, pp. 213–216.

Sperber, Dan, Fabrice Clement, Christophe Heintz, Olivier Mascaro, Hugo Mercier, Gloria Origgi, and Deirdre Wilson (2010). "Epistemic vigilance". In: *Mind & language* 25.4, pp. 359–393.

Stantcheva, Stefanie (2023). "How to run surveys: A guide to creating your own identifying variation and revealing the invisible". In: *Annual Review of Economics* 15, pp. 205–234.

Stettler, Dan D and Richard Axel (2009). "Representations of odor in the piriform cortex". In: *Neuron* 63.6, pp. 854–864.

Stewart, Neil, Nick Chater, and Gordon DA Brown (2006). "Decision by sampling". In: *Cognitive psychology* 53.1, pp. 1–26.

Sturgis, Patrick and Rebekah Luff (2020). "The demise of the survey? A research note on trends in the use of survey data in the social sciences, 1939 to 2015". In: *International Journal of Social Research Methodology*, pp. 1–6.

Tourangeau, Roger (2021). "Survey reliability: models, methods, and findings". In: *Journal of Survey Statistics and Methodology* 9.5, pp. 961–991.

Triantafyllou, Christina, Richard D Hoge, Gunnar Krueger, Christopher J Wiggins, Andreas Potthast, Graham C Wiggins, and Lawrence L Wald (2005). "Comparison of physiological noise at 1.5 T, 3 T and 7 T and optimization of fMRI acquisition parameters". In: *Neuroimage* 26.1, pp. 243–250.

VanderWeele, Tyler J (2011). "Principal stratification–uses and limitations". In: *The international journal of biostatistics* 7.1.

Vannette, David L and Jon A Krosnick (2017). *The Palgrave handbook of survey research*. Springer.

Varaine, Simon (2023). "How dropping subjects who failed manipulation checks can bias your results: An illustrative case". In: *Journal of Experimental Political Science* 10.2, pp. 299–305.

Vasudevan, Varsha, Aditya Murthy, and Radhakant Padhi (2024). "Modeling kinematic variability reveals displacement and velocity based dual control of saccadic eye movements". In: *Experimental Brain Research* 242.9, pp. 2159–2176.

Vulkan, Nir (2000). "An economist's perspective on probability matching". In: *Journal of economic surveys* 14.1, pp. 101–118.

Walker, Heather EK and Lana M Trick (2018). "Mind-wandering while driving: The impact of fatigue, task length, and sustained attention abilities". In: *Transportation research part F: traffic psychology and behaviour* 59, pp. 81–97.

Weinstein, Yana (2018). "Mind-wandering, how do I measure thee with probes? Let me count the ways". In: *Behavior research methods* 50, pp. 642–661.

Westwood, Sean J, Justin Grimmer, Matthew Tyler, and Clayton Nall (2022). "Current research overstates American support for political violence". In: *Proceedings of the National Academy of Sciences* 119.12, e2116870119.

Wilson, Timothy D and Nancy Brekke (1994). "Mental contamination and mental correction: unwanted influences on judgments and evaluations." In: *Psychological bulletin* 116.1, p. 117.

Wolff, Irenaeus (2019). "The reliability of questionnaires in laboratory experiments: What can we do?" In: *Journal of Economic Psychology* 74, p. 102197.

Wong, Yi-Sheng, Adrian R Willoughby, and Liana Machado (2023). "Reconceptualizing mind wandering from a switching perspective". In: *Psychological research* 87.2, pp. 357–372.

Zaller, John and Stanley Feldman (1992). "A simple theory of the survey response: Answering questions versus revealing preferences". In: *American journal of political science*, pp. 579–616.