Essays on the Environmental and Behavioral Determinants of Health Outcomes

by

Kelly Hyde

B.S. in Economic Analysis, Binghamton University, 2015

B.A. in English Rhetoric, Binghamton University, 2015

B.A. in Mathematics, University at Buffalo - State University of New York, 2016

M.A. in Economics, University of Pittsburgh, 2018

Submitted to the Graduate Faculty of

the Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2022

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Kelly Hyde

It was defended on

March 24, 2022

and approved by

Osea Giuntella, Assistant Professor of Economics, University of Pittsburgh

David Huffman, Professor of Economics, University of Pittsburgh

Randall Walsh, Professor of Economics, University of Pittsburgh

Karen Clay, Professor of Economics and Public Policy, Carnegie Mellon University

Andrea La Nauze, Lecturer of Economics, University of Queensland

Essays on the Environmental and Behavioral Determinants of Health Outcomes

Kelly Hyde, PhD

University of Pittsburgh, 2022

This dissertation consists of three essays on environmental, behavioral, and health economics. Chapter 1 presents an experimental study of how individuals use group-labeled disaggregated information to form subjective beliefs about their own prospects under uncertainty. This study finds consistent evidence of a "category specificity heuristic": individuals assume that information about others who share observable characteristics with them, such as gender, age, race, or educational attainment, is more informative than information about others who do not, even in cases where there is no plausible causal relationship between group membership and outcomes. Implications for individuals' beliefs about health risks are discussed. Chapter 2 presents an observational study of the relationship between potable water availability and heat-related mortality in South Africa which demonstrates that increased potable water supply is an effective community-level adaptation to excess heat. The findings of this study suggest that investments to increase the reliability and accessibility of potable water, especially in developing contexts, may ameliorate the long-run mortality consequences of climate change. Chapter 3 presents a study which establishes a causal link between water quality violations and food insecurity in the United States. This study finds that during an active water quality violation in their county of residence, the nutritional content of lower-income households' grocery store purchases differentially declines relative to wealthier households in the same county, with effect sizes that increase with the duration of the violation.

Table of Contents

List of Tables

List of Figures

Preface

There were hundreds of moments I thought this dissertation would never exist. That it does is a testament to the people who supported me in those moments: those who walked with me to get coffee (or, in pandemic times, hopped on Zoom) and talked me down from my latest crisis; those who reassured me my project would still contribute to the literature despite the latest worryingly similar working paper I found; those who never lost patience helping me practice my "elevator pitch" over, and over, and over again; and those who had very little interest in the world of academic economics research but nonetheless listened to my problems. I owe these people more than they know. Among them are my partner Brian, who loved me even at my most anxious and burnt out; my mother Sheila, who always helped me maintain perspective; my classmates Bea Ahumada, Mallory Avery, Neeraja Gupta, and Marissa Lepper, who became family; and my best friends Danielle and Rosalie, who likely learned *much* more about the inner workings of an economics PhD program than they ever cared to know, but never complained. I am also thankful for the numerous mentors who helped these chapters and the start of my career take shape: my dissertation committee; my fellowship supervisor Dr. Ilia Murtazashvili who provided invaluable support toward the end of my program; and my AMIE mentors Dr. Teevrat Garg and Dr. Kathleen Segerson. Finally, I honor the memory of Dr. Werner Troesken, whose zeal for research and wellspring of encouragement for graduate students set an invaluable example during my short but meaningful time as his advisee.

1.0 Learning About Subjective Uncertainty: Overinference from Observable Characteristics in Disaggregated Data

Individuals facing uncertainty frequently use information at varying levels of (dis)aggregation about others' realized outcomes in similar environments to form subjective beliefs about their own prospects. The use of disaggregated information introduces another level of subjectivity: the individual's beliefs about the informativeness of signals based on the category to which they pertain. In this paper, I experimentally study individual belief updating in contexts where outcomes do not meaningfully differ across categories. I find consistent evidence of a bias: individuals incorrectly assume that information about someone in a particular category is more informative about the prospects of others in the same category than about others in different categories, even when they directly observe the underlying process that assigns individual prospects. As a result, when they receive noisy disaggregated information, their posterior frequently features exaggerated differences across categories, especially when the information reinforces a preexisting misperception in their prior. When these incorrect beliefs pertain to their own category, individuals subsequently acting on these beliefs take on a different level of risk than their risk preferences imply they would if their beliefs had been correct. These results suggest that providers of information that may influence individuals' risky behaviors should carefully select a level of (dis)aggregation which balances personal relevance with statistical precision.

1.1 Introduction

Many real-world decisions involving uncertainty require the decision maker to base their choices on a subjectively perceived likelihood of each possible outcome. These decisions arise in a variety of contexts and are often consequential: an undergraduate's choice of major is influenced by their beliefs about the most likely subsequent employment and salary outcomes; a home buyer's choice of neighborhood is influenced by their belief about the likelihood of being a victim of a crime or a natural disaster; and an individual's choice of whether to wear a mask in public during a pandemic is influenced by their belief about their risk of exposure to disease. In all of these cases, the individual cannot observe their own true probability distribution over outcomes prior to making a decision. Instead, they must form a belief about this probability distribution based on the information available to them, including the observable outcomes of others who have already gone through a similar situation. The formation and validity of subjective probabilities have long been of interest to economists and psychologists (e.g., Bassett and Lumsdaine 2001, Cerroni 2020, Chiodo 2004, Epstein and Zhang 2001, Hurd 2009, Hurd and McGarry 2002, Kahneman and Tversky 1972, Kieren and Weber 2021, Tversky and Kahneman 1974).

When investigating others' realized outcomes in the same decision environment to guide their beliefs about their own prospects, the decision maker may encounter information at varying levels of (dis)aggregation, with categories frequently defined by either demographics, geographic location, or other identifying characteristics. The relative value of aggregated and disaggregated information with a fixed sample size is based on a tradeoff between bias and variance; disaggregating statistics explicitly accounts for heterogeneity across categories (i.e., reducing omitted variable bias), but at the cost of increased variance when each category is a proper subset of the full sample. Thus to effectively incorporate disaggregated information into their beliefs about their own prospects, the decision maker must have well-calibrated beliefs about this tradeoff, weighing the gain in personal relevance from focusing on the outcomes of those in the same category (i.e., the extent to which membership in a particular category predicts one's outcomes) against the increased likelihood of observing signals skewed by noise.

In this paper, I conduct a tightly controlled lab experiment and companion survey on health outcomes which demonstrate that individuals have biased beliefs about the informativeness of disaggregated information. In particular, when outcomes *do not* significantly differ across the categories by which information is disaggregated, individuals nonetheless appear to believe that information about a particular category is highly informative about the prospects of individuals in that category and significantly less (or, in some cases, not at all) informative about the prospects of individuals in other categories. Remarkably, this is true even when individuals directly observe the underlying process which generates individual prospects prior to receiving disaggregated information, suggesting that this way of thinking is a heuristic which individuals apply whether or not there is a plausible reason to believe outcomes differ across categories. As a result, observing noisy disaggregated information frequently results in posterior beliefs which exaggerate the differences in outcomes between categories, especially when the information pertains to the individual's own category and/or reinforces a preexisting error in the individual's prior. When this leads the individual to have incorrect beliefs about their own prospects, they subsequently take on a different level of risk than they likely would have if their beliefs had been correct.

The results of this paper are informative for economic theory, as they suggest a benefit of modeling an additional layer of subjective beliefs that is not typically considered in models of decision making under uncertainty: their beliefs about the (relative) informativeness of information about others. When using disaggregated information about others to form beliefs about their own prospects, the decision maker applies a subjective belief about the correlations between categories; each subjective correlation implies a weight the individual should apply to information about others in that category. Prior literature on correlation neglect (Ellis and Piccione 2017, Enke and Zimmermann 2019, Kallir and Sonsino 2009, Levy and Razin 2015a,b) suggests individuals frequently underestimate or ignore such correlations. However, the results of this paper suggest errors in correlation perception in both directions: individuals underestimate correlations *across* categories and overestimate correlations *within* categories. In other words, individuals appear to believe they are more different from others in different categories and more similar to others in the same category than they actually are. This finding is consistent with models of object categorization from psychology. In particular, Gestalt theory (Ellis 1938, Wertheimer 1938) posits that individuals tend to group together objects that appear similar in some observable characteristic, and contemporary experiments in sensory perception have found evidence in favor of this hypothesis in the visual, auditory, and tactile domains (Gallace and Spence 2011, Wagemans et al. 2012). Theories of social categorization in social psychology (Allport 1924, Campbell 1958, Krueger and DiDonato 2008, Tajfel et al. 1971, Turner et al. 1987) similarly hypothesize that individuals tend to place other individuals in groups based on shared social identities. Using the language of this literature, the results of this paper suggest a perception of false consensus among outcomes within a category and false uniqueness across categories (Mullen et al. 1992, Perloff and Brickman 1982, Ross et al. 1977, Suls and Wan 1987).

The findings of the paper are also relevant for policy, suggesting additional, psychological factors to be considered by policy makers when choosing how to provide information. Providing disaggregated information may seem to offer a benefit in terms of making this more personally relevant for decision makers, with little downside if individuals are assumed to correctly account for the additional noise that comes from disaggregated information. Indeed, potential benefits of the increased personal relevance of information about others who share an observable characteristic with the decision maker have been demonstrated in recent research on "nudges" involving disaggregated information, such as notifying individuals about their energy consumption relative to their neighbors to encourage conservation (Allcott and Rogers 2014), intentionally featuring spokespeople who openly share characteristics with the target population in a public health campaign (Keene et al. 2021), and using information about fellow students to correct misperceptions in the frequency of alcohol consumption among college students (Neighbors et al. 2004, 2006). If, instead, people strongly underestimate noise, and furthermore systematically overestimate the extent to which variation across different categories reflects differences in common unobserved risk, then disaggregated information may be detrimental. This is especially true because the results of this paper suggest individuals update their beliefs based on these misperceptions even in the ideal scenario where the underlying process which assigns individual prospects is explained to them prior to observing the information. In most real-world scenarios, this data-generating process is at best imperfectly observable, meaning there is even less discouragement of over-inferring from observable characteristics in disaggregated data.

The experiment requires participants to guess an unobserved parameter X which has been randomly assigned to them, and subsequently choose between a range of possible sure payments and entry into a lottery with X chance of winning \$5. Participants are fully aware of how X is determined, but they do not observe their assigned X at any point in the experiment. In this way, the experiment emulates real-world risky environments in which the exact probability of each outcome is unknown. It also deliberately creates an environment in which there is no rational reason for participants to believe that prospects are significantly correlated within-category.

Prior to making their guess and decision, participants are shown the outcomes (win or loss) of six "test runs" of the lottery using the respective assigned X values for six other individuals in the session, three in-group (defined in this context as individuals who share their gender identity) and three out-group. Because X is randomly assigned, this signal has zero informational content and should be disregarded. Despite this, prior literature suggests many will update their beliefs based on it (Chadd et al. 2021, Flepp 2021, Kieren and Weber 2021, Nimark and Sundaresan 2019). Conditional on treating the signal as informative, it is in participants' best interest to ignore the group labeling; segregating the information by group decreases the effective sample size (thereby increasing the influence of noise) without increasing the relevance of the information.[1]

The results suggest a systematic overestimation of in-group correlations and underestimation of correlations between groups. While observing an additional in-group test run win increases an individual's guess of their own X by 10 percentage points and their certainty equivalent of the lottery (i.e., price list switching point) by \$0.22 on average, observing an additional out-group test run win has no significant effect on either outcome. The response to in-group information is so strong that the number of in-group test run wins observed is

[1] An implicit assumption here which is true in the experimental design is that individuals regardless of group are assigned their prospects from the same distribution, meaning that an in-group signal contains exactly the same amount of information about the underlying distribution as an out-group signal.

statistically a stronger predictor of a participant's certainty equivalent of the lottery than their self-reported general risk tolerance. In summary, individuals' beliefs about their own prospects exclusively respond to in-group information, and this belief updating leads them to take on a different level of risk in subsequent choices.

The results also suggest the individuals most likely to misperceive the relative informativeness of disaggregated information are those who identify most strongly with their own category. In particular, inspired by Chen and Li (2009) and Tajfel et al. (1971), participants in the experiment complete an allocation task between a pair of receivers, one in-group and one out-group, and are asked to self-report how closely attached they felt to their group throughout the experiment. The beliefs of individuals who allocated more to the in-group receiver and/or reported a high degree of attachment to their group during the experiment responded more strongly to the number of observed in-group wins, while still not responding to out-group wins. This suggests a potential intuitive mechanism for the observed results: individuals who identify more strongly with their group are more likely to believe that, for a particular characteristic, within-category correlations are high (i.e., a generalized "false consensus" as defined by Ross et al. (1977)) while cross-category correlations are low (i.e., a generalized "false uniqueness" (Snyder and Shenkel 1978, Suls and Wan 1987)). Given this belief, it is reasonable for an individual to respond strongly to observed in-group outcomes while disregarding out-group outcomes.

To supplement the experimental results, I conduct an online survey which asks participants to guess the respective prevalences of health conditions among demographic groups of individuals in the United States. Like the experiment, the survey elicits how individuals update their beliefs after observing noisy disaggregated signals in an environment where the true expected value is identical for both groups. This complements the experiment in multiple ways: it includes a broader range of categories, enabling the assessment of heterogeneity across possible types of categorization; it pertains to information about which participants are more likely to have meaningful priors; and it directly tests whether the results of the experiment generalize to the motivating example of health-related beliefs.

Similar to the experiment, the survey results demonstrate that noisy disaggregated information leads individuals to believe that outcomes differ significantly across social categories when they do not. Across four pairs of health conditions and population groups, after receiving information about a particular category, individuals' beliefs moved between two and ten times more for that category than for the other category. Moreover, I find suggestive evidence that, at least in some cases, the degree of asymmetric updating is larger among individuals who received information about their own category. In other words, while individuals appear to underestimate cross-category correlations in general, they are especially likely to underestimate the correlation between their own category and another category.

This paper contributes to the literature on the influence of categorizations, such as group identity, on individual beliefs and behavior. Social scientists have extensively studied group-related perceptions and behaviors, including distorted beliefs about ability and performance (Cacault and Grieder 2019, Casoria et al. 2020, Paetzel and Sausgruber 2018, Sandberg 2018), in-group favoritism (Ben-Ner et al. 2009, Chen and Li 2009, Grimm et al. 2017, Güth et al. 2009, Ockenfels and Werner 2014), and (mis)trust and cooperation (Ahern et al. 2014, Arbatlı et al. 2020, Delavande and Zafar 2015, Goette et al. 2006, Meier et al. 2016). This paper applies the consistent finding of bias in favor of one's own group to an information processing domain, hypothesizing that the assumptions which lead individuals to favor their own group over others may also encourage them to regard information about other groups as irrelevant to themselves, even when it is equally or even more informative. This hypothesis is also closely related to the literature on peer effects, especially in the risk-taking domain (Bougheas et al. 2013, Gioia 2019, Lahno and Serra-Garcia 2015, Sontuoso et al. 2021). While this literature primarily focuses on the influence of information about others' *choices* on an individual's beliefs and choices in the same domain, this paper focuses on a context in which the information is about others' *outcomes* and the choices that led to those outcomes are unobserved. This approach more closely represents contexts in which the primary information available is anonymous statistics, such as determining your risk of exposure to an infectious disease or predicting your salary after graduating with a particular degree.

The findings of this paper are also complementary to the literature on errors in information processing. Prior literature has demonstrated that individuals update their beliefs based on information which would optimally be ignored (Chadd et al. 2021, Flepp 2021, Kieren and Weber 2021), misperceive correlations between signals and risks (Ellis and Piccione 2017, Enke and Zimmermann 2019, Kallir and Sonsino 2009, Levy and Razin 2015a,b), and disregard or selectively interpret sample bias and noise (Enke 2020, Hamill et al. 1980, Nimark and Sundaresan 2019, Tversky and Kahneman 1974). This paper studies the effect of these errors on disaggregated information. Each plays a role in the findings of the experiment and survey: participants treat noisy, uninformative information as though it is informative, neglect cross-group correlations, and treat extremely small samples of information about a group as diagnostic.

The rest of the paper proceeds as follows. Section 2 discusses the design of the experiment. Section 3 summarizes the results of the experiment. Section 4 describes the supplemental survey design. Section 5 reports the results of the survey. Section 6 concludes.

1.2 Design of Lab Experiment

A total of 232 undergraduate and graduate student participants were recruited from the Pittsburgh Experimental Economics Laboratory (PEEL) at University of Pittsburgh in June 2020. By necessity because of the closure of in-person facilities during the COVID-19 pandemic, sessions were conducted virtually over Zoom. Participants progressed through a series of computerized tasks coded in oTree (Chen et al. 2016). Each session, including check-in and initial instructions, took 45-60 minutes. The experiment was preregistered with the AEA RCT Registry under trial number AEARCTR-0007830.

At the very beginning of the experiment, participants are asked about their demographics, including gender and age. This is done at the beginning deliberately, because participants are subsequently placed in groups based on shared gender identity. Additionally, participants are asked to self-report their risk preferences on a 10-point scale using the question validated by Dohmen et al. (2011).

Prior to the decision stage, participants were told that they had been placed in a group with 3 other participants of the same gender. This was done for both practical and investigative reasons. Gender as a grouping was well suited to the needs of the experiment, which required exact balance between two groups formed based on shared characteristics. While the hypothesis tested by the experiment is not specific to gender and intended to be more general, gender has also been a grouping of particular interest in related literature, which has documented systematic differences in risk perception between men and women (Finucane et al. 2000, Flynn et al. 1994, Palmer 2003). Thus groups based on gender were both convenient and of special interest. To avoid confounding effects from overlapping groupings, participants were not given any information about their group-mates other than their gender, and every session contained a minimum of 8 participants of each gender so that participants could not discern who was in their group by looking at others' video on Zoom. Additionally, some participants were randomly assigned to one of two priming conditions. Because these priming conditions did not significantly influence individual behavior in the experiment and the key results hold in the subsample that did not receive any priming, I pool the data in the analyses presented in this paper.[2]

1.2.1 Decision Stage

In the decision stage of the experiment, participants are asked to guess an unobserved, randomly assigned parameter X, and to choose a switching point in a price list with gradually increasing sure payments and a fixed lottery with X chance of winning \$5. Guesses are incentivized using the binarized scoring rule (Hossain and Okui 2013), while the switching point is incentivized by implementing their choice in a randomly selected row of the price list. Following the recommendation of Danz et al. (2020), the binarized scoring rule was not explained in detail to participants; instead, participants were simply told that their

[2]For more information about the priming conditions, see Appendix Section A.1.

chance of earning the bonus was based on how close their answer was to the correct answer, and therefore it was in their best interest to report their true best guess. Participants are informed that X is individually randomly assigned by drawing a number from the uniform distribution over integers between 0 and 100. However, participants observe neither their own assigned X nor anyone else's at any point in the experiment.

Prior to making their guess and choice, participants are shown the outcomes of six group-labeled "test runs" of the lottery. Each test run uses the actual assigned X of another participant in the session: the participant's three group-mates, and three randomly selected individuals of a different gender. A visual example of actual information a participant saw in the experiment is provided in Table 1. To rule out salience of left-column information as a confound, the order of columns was randomized by individual so that half of the participants saw their own group on the left and the other half saw the other group on the left.

Since X is randomly assigned through independent draws from a uniform distribution, this information has zero informational content pertaining to the participant's own X and should be disregarded. Since the participant does not observe any draws based on their own X, they have no information beyond the fact that their X was drawn from a uniform distribution over integers between 0 and 100. Therefore the optimal guess is the expected value of the distribution, i.e., $X = 50$. Even if the participant erroneously interprets the information as informative about their own X (i.e., believes that X is correlated across participants), the information comprises a very weak, noisy signal about others' assigned X; therefore even guesses of individuals who fail to ignore the information should be close to $X = 50$.

Immediately after seeing the information, participants are asked to guess their own X on a slider between 0 and 100. After submitting this guess, participants are asked to choose a switching point between Option A and Option B in the price list depicted in Table 2. In other words, participants are asked to click the smallest value of Option B (the sure payment) they would prefer over Option A (the fixed lottery with X chance of winning \$5).

After guessing their own assigned X and choosing a switching point in the price list, participants allocate \$1 on a slider in increments of \$0.01 between two other participants in the session: one of their group mates (and thus of the same gender), and one randomly selected individual of a different gender. As in Tajfel et al. (1971) and Chen and Li (2009), this allocation task is intended to measure an individual's degree of in-group favoritism. Participants are not given any additional information about their assigned receivers aside from their gender and group membership. To eliminate concerns of retribution, participants are also reassured that their receivers will not know their identity.

This serves as an established group bias from prior literature which is plausibly correlated with selective attention to in-group information. While there are several potential reasons for allocating a larger share to an in-group receiver, including social preferences, desire to follow norms, or application of a simple heuristic, it is plausible that an individual who is influenced by group labels in an allocation setting is also influenced by them in other decision environments in a similar way. For example, if people allocate more to in-group receivers because they assume individuals in the same group are similar to them, this way of thinking would likely also influence their perception of the relevance of information about others to themselves based on their group. Alternatively, if favoring an in-group receiver in an allocation is a heuristic as argued by Guala and Filippin (2017), applying essentially the same heuristic (in the absence of other information, look for the one in my group) to group-labeled information would imply selective attention to in-group information.

At the end of the experiment, participants complete a short survey about their experience in the study. As a standard test of experimenter demand effects, participants are asked what they thought the experiment was about in a free-response question. Inspired by the exit survey in Chen and Li (2009), participants were also asked how closely attached they felt to their group throughout the experiment on a scale from 1 to 10.

1.3 Results of Lab Experiment

Figure 1 graphically depicts the main result of the experiment: participants' average guesses about their assigned X respond strongly and linearly to in-group test run wins, while they do not respond at all to out-group test run wins. Each additional in-group test run win observed increases the average guess of X by about 10 percentage points. Each pairwise difference in Figure 1a is statistically significant ($p \sim 0.015$ for zero wins vs one win, and $p < 0.01$ for the rest), while the pairwise differences in Figure 1b are insignificant and vary in sign.

Table 3 confirms the findings in Figure 1 and extends them to the subsequent price list switch point. Since each row of the price list increased the sure payment by \$0.50, the coefficient in the second column of Table 3 can be approximately interpreted as an average increase of \$0.22 in the certainty equivalent of the lottery with X chance of winning \$5 for each additional in-group test run win observed. While the Dohmen et al. (2011) measure of self-reported risk tolerance on a 10-point scale was still positively associated with the switch point, the number of in-group wins observed was a much stronger predictor, suggesting that the information caused some individuals to make a choice they would not have if they had accurate beliefs about X. To rule out salience of left-column information as a mechanism of the effect, columns 3 and 4 add an interaction term for both in-group and out-group wins observed with an indicator for the participant seeing their own group's information on the left; the interaction terms are insignificant in all cases.

In summary, Figure 1 and Table 3 demonstrate that when provided group-labeled information about others' outcomes in a risky environment, even in the absence of a plausible reason to believe individual outcomes are predicted by group membership, individuals' beliefs respond strongly to in-group information while ignoring out-group information. As a consequence, when the in-group information is not representative of the expected value, they end up with incorrect beliefs. It is important to acknowledge that in this environment, *any* updating based on the information is likely to lead to incorrect beliefs, since others' outcomes are entirely irrelevant to an individual's prospects. However, conditional on failing to recognize the irrelevance of the information, individuals would have been better off on average updating based on the proportion of wins among all six test runs observed, regardless of group. About 30% of participants (69 out of 232) observed 3 total wins out of 6 test run results. Assuming the participant's guess of their own X is based on the proportion of wins they observed, this would coincide with $X = 50$. Instead, among the 3 in-group test run wins participants actually took into account, observing 50% wins was not possible,[3] meaning that all possible signals were not representative of the true expected value.

1.3.1 Heterogeneity and Potential Mechanisms

The results in Table 4 reflect two significant moderators of the effect of in-group information. Column 1 suggests that the effect is strongly positively related to the amount participants allocated to the in-group receiver. In other words, the more individuals allocated to their group-mate, the more their X guess responded to observed in-group wins in the decision stage, and vice-versa. Because all participants completed their allocation after seeing the test runs and guessing their X, it is not technically possible to conclusively determine the direction of causality in this relationship.[4] However, the notion that observing more in-group wins led participants to allocate more to the in-group receiver is unintuitive; if anything, one would expect the opposite, as individuals who saw their group-mates losing in the decision stage might have used the allocation as an opportunity to compensate for this if they are inequality averse or concerned with fairness (Cappelen et al. 2016, 2017, Charness and Rabin 2002, Fehr and Schmidt 1999). It is more intuitively likely that individuals who came in to

[3]Since participants observed 3 in-group wins, the possible numbers (percentages) of wins they incorporated into their guess of their own X were 0 (0%), 1 (33%), 2 (67%), or 3 (100%). None of these outcomes coincide with the expected value of 50%. Assuming individuals begin with a prior of $X = 50$, this means that posterior guesses after observing test runs will always trend away from the true expected value.

[4]This was purposeful to avoid generating experimenter demand effects. If participants completed the in-group out-group allocation before seeing the group-labeled information, it may have become obvious that the experiment was about in-group bias, thereby confounding the main result. While allocations may have been influenced by test run information, given the central focus of the paper, this is preferable to the other way around.

the experiment with a stronger pre-existing group bias, reflected in their allocation, were in turn more likely to believe in-group information was relevant to them while out-group information was not. The results in column 2 of Table 4 reflect a similar, albeit weaker, positive relationship between the response to in-group information and self-reported attachment to one's own group during the experiment. As one would expect, the two measures of group identification used in columns 1 and 2 respectively are positively correlated ($\rho \sim 0.4$), so it is unsurprising that both interaction effects go in the same direction. Similarly, in both cases, there is a null effect of group identification on the response to out-group information, ruling out the possibility that stronger group identifiers simply paid more attention to information in general.

1.3.2 Session-Level Beliefs

In the exit survey, participants were asked about the average assigned X among men and women respectively on a 5-point Likert scale ranging from "Women's average X was much higher than men's" to "Men's average X was much higher than women's." These session-level beliefs complement participants' guesses of their own X in two ways. First, they shed light on whether participants apply their belief about their own X to others in the same broader category, not just those in the same constructed group.[5] Second, the session-level beliefs provide a way to determine whether or not out-group information was maintained in participants' working memory during the experiment. The ignoring of out-group information demonstrated in Figure 1 and Table 3 could be consistent with two possible stories: either participants completely ignored out-group information, perhaps not even bothering to look at it; or participants did acknowledge out-group information, but consciously decided it was not relevant to their own X.

Figure 2 and Table 5 show the responses of session-level beliefs to the test run wins

[5] Technically, there were two layers of "groups" in the experiment: the broader categories of "men" and "women" respectively, and then the specific subgroups of 4 into which participants were placed, which were nested within the broader gender-based groups. Because of this, the session-level beliefs are necessary to rule out the possibility that individuals considered the information about *their specific group of four* as diagnostic of their own prospects, but not reflective of their gender's prospects as a whole.

observed in the decision stage. Participants' beliefs linearly respond to the difference in wins between men and women, respectively, in the appropriate direction. Across all possible information observed, there is a slight bias toward men; the average belief among individuals who observed the same number of wins among men and women is significantly different from 3 (i.e., "Men and women had roughly the same average X" on the scale), and among participants who observed one *less* win among men than among women, the average belief is almost exactly 3. The degree of bias toward men was not significantly different between men and women, as demonstrated by the coefficient on the male indicator in Table 5.

Column 2 of Table 5 adds an interaction term for the participant's own gender. The coefficient, which is statistically significant, suggests that women's session-level beliefs responded about 64% more strongly to the disparities observed in the test runs than men's. To differentiate between positive and negative belief updating, Appendix Table 16 reports separate regressions for optimistic and pessimistic beliefs about the average X among a participant's own gender respectively. These results suggest that women were significantly more likely to believe specifically that the average X among women was lower than the average X among men if they had observed men outperforming women in the test runs (i.e., negative belief updating). This provides an interesting contrast to the results in Appendix Table 15, which suggest there was no significant difference between men and women in belief updating about an participant's own X or subsequent choices based on the test run information. Taken together, these results provide suggestive evidence that while men and women's respective belief updating patterns about their own *individual* prospects are similar, their beliefs about their group as a whole may diverge. In particular, women may be more likely to interpret "bad news" for themselves as an indication that women in general are or will be worse off in that context, while men may be more likely to believe that bad news for themselves reflects an idiosyncratic anomaly.

1.3.3 Assessing Experimenter Demand

While recent studies have suggested that experimenter demand effects are modest even when the researcher explicitly signals their hypothesis in the instructions (de Quidt et al. 2018, Mummolo and Peterson 2019), it remains a potential concern for this study. Multiple features of the experimental design were intended to mitigate this concern. First, participants' guess of their own X and subsequent price list decisions were incentivized, making the distortion of answers to meet the researcher's perceived expectations financially costly. In the explanation of these incentives, participants were explicitly told that it was in their best interest to report their true best guess. Second, the randomized priming conditions[6] served the additional function of muddling the purpose of the experiment, especially because the priming stages took substantially more time than the decision stage. As discussed in Appendix Section A.1, the priming appears to have made individuals think about the intended types of group identification; however, the estimated effects were not significantly different across treatments. Therefore participants who had demonstrably different beliefs about the purpose of the experiment exhibited similar responses to in-group information.

Experimenter demand is also unlikely to explain the results in Table 4 on group identification as a moderator. The effect of in-group information on individuals' beliefs about their own X is almost entirely moderated by the amount allocated to the in-group receiver. As a result, for experimenter demand to explain the effect, participants must have simultaneously anticipated that the experimenter wanted them to pay attention to in-group test run wins *and* wanted them to allocate more to their in-group receiver. While this is technically possible, a much more direct explanation is that strength of group identification simultaneously influences in-group/out-group allocations and belief updating responses to group-labeled information.

Finally, in the exit survey, participants were asked (and required to answer) what they thought the experiment was about in an open response question. In addition to assessing the efficacy of the priming conditions, this provides a way to determine whether participants

[6] For an explanation of the priming conditions, see Appendix Section A.1.

could correctly identify the objective of the experiment. The ideal scenario is that participants thought the experiment was about something unrelated to the primary outcome (the guess of their own X) such as gender differences/fairness, teamwork in groups, or risk preferences. In total, 146 participants (63% of the sample) mentioned either the word *gender* or a related word, such as *men*, *women*, *male*, or *female*; 23 (10%) mentioned *group* or *team*; 61 (26%) mentioned *risk*, and 41 (18%) mentioned the word *bias*. Among those who said *bias*—perhaps the closest to the research question—many referred to either the in-group/out-group allocation or one of the priming conditions. Only one participant explicitly referred to the test runs, while 32 (14%) mentioned the allocation task. Two participants said *guess*; however, both of these participants were referring to a priming condition as opposed to the X guess. Overall, participants' answers suggest that while they understood the purpose of the priming conditions and allocation task, they could not identify the purpose of the main decision stage beyond simply measuring "risk preferences," which was the intended outcome.

1.4 Design of Supplemental Survey

To supplement the lab experiment described and analyzed in the previous sections, I designed a survey to be run on the online study recruitment platform Prolific in September 2021. In total, 202 participants completed the survey. The survey was coded in Qualtrics and took an average of 13 minutes to complete. For their participation, survey respondents were paid a guaranteed \$2 and given the opportunity to earn a \$1 bonus based on their answers. The bonus was incentivized using the binarized scoring rule (Hossain and Okui 2013) for their answer to one randomly selected part of one question. Like the experiment, following the recommendation of Danz et al. (2020), the binarized scoring rule was not explained in detail to participants; instead, participants were simply told that their chance of earning the bonus was based on how close their answer was to the correct answer, and therefore it was in their best interest to report their true best guess for every question.

The survey featured 4 two-part questions in which participants guessed the prevalence of particular health conditions among two demographic groups of Americans. The correct answers were based on the 2019 National Health Interview Survey. The questions pertained to the following pairs of health conditions and demographic groups: the respective percentages of men and women who have ever been diagnosed with any type of cancer, the respective percentages of White and Asian Americans who have ever been diagnosed with any type of diabetes, the respective percentages of high school graduates and undergraduate degree holders who have ever been diagnosed with anxiety, and the respective percentages of 18-29 and 30-39 year olds who have ever been diagnosed with depression. These questions were presented to participants in a random order.

Participants' guesses in each question were given in the form of a 3x3 matrix which represented a binned version of their perceived joint probability distribution of health condition prevalences by demographic group. Figure 3 provides an example taken directly from the survey of this joint probability distribution matrix. Participants were told that one of the cells labeled A through I contained the correct answer, and were asked to enter their perceived probability (as a number between 0 and 100) of each particular cell containing the correct answer.

The pairs of health conditions and demographic groups were deliberately selected so that the true prevalence in the 2019 NHIS is nearly identical between the two groups, and the correct answer is always contained in Cell E (both prevalences are between 9% and 16% when rounded to the nearest whole number). This is done to mirror the environment in the experiment in which the true expected value is the same for both groups. As a result, in every question, the "correct answer" places a probability of 0 in all cells except Cell E, which is assigned a probability of 100.[7]

In each question, participants filled in this joint probability distribution matrix twice: once as a prior, and once as a posterior after seeing the prevalence of the health condition

[7] For simplicity, this assumes the 2019 NHIS is perfectly representative of the population. In the survey, the question was worded in such a way that it was clear participants were guessing the percentages according to the survey, as opposed to the true population percentages.

derived from a subsample of 50 randomly selected individuals in one of the groups. The process for building this information was to draw ten random subsamples of 50 individuals in each demographic group from the NHIS with replacement and calculate the prevalence of the associated health condition for each subsample.[8] Then, among these ten subsample prevalences, the minimum and maximum were selected as information to be shown in the survey. Thus while this information was truly drawn from a random subsample of 50 individuals in that group, the information participants were shown was deliberately different from the true group mean. The outcome of interest is the difference between these posterior and prior beliefs.

Within each question, participants randomly saw either "low" (a subsample which is significantly lower than the true group prevalence) or "high" information. Similar to the experiment, the information given has low informational content.[9] Information was given about women, white Americans, 18-29 year olds, and undergraduate degree holders; in all cases, a random sample of 50 comprised approximately 1 percent or less of the total number of NHIS respondents in that group. As a result, the prevalence among a random sample of 50 is a noisy signal of the true group prevalence.

[8]Participants in the survey were asked to guess the population-level prevalence of a condition based on a nationally representative sample of Americans. Thus these calculations took survey sampling weights into account because the correct answers were based on the weighted proportion of individuals reporting a particular diagnosis in the survey sample. This is why, for example, participants could be shown a subsample prevalence of 23% of cancer among women even though a raw proportion of 23% would not be possible in a sample of 50 indivisible units.

[9]Unlike the experiment, the informational content is not zero. To calculate the probability that the prevalence among a random subsample of 50 NHIS respondents in a particular group reflected the correct answer, 10,000 random samples were drawn with replacement. The percentage of samples with a prevalence between 9% and 16% (the correct answer in all cases) was 51% for cancer among women, 41% for diabetes among white Americans, 52% for anxiety among college graduates, and 48% for depression among 18-29 year olds. In all cases, this percentage is greater than random chance (1/3).

1.5 Results of Supplemental Survey

Table 6 reports the demographics recruited from Prolific for the supplemental survey. The sample skews more female, younger, and more educated than the general population. However, critically, the survey sample is closer to balanced in terms of "in-group" and "out-group" relative to the information given in each question. The groups about which information was provided were deliberately selected to be the anticipated largest groups in the study sample so that comparisons of in-group versus out-group responses to information would have sufficient statistical power. This was achieved across all types of demographic groups. However, one concern with the sample which is not represented in Table 6 is that the age-by-gender distributions are not balanced. Women in the sample were significantly younger on average than men in the sample. Since the main results of the survey are within-subject difference-in-difference comparisons of priors and posteriors by group, this should not be a concern for internal validity. However, it does potentially confound between-subject in-group versus out-group updating comparisons, and may limit the generalizability of the findings if the belief updating patterns of young women are distinct from other population groups.

Tables 7 and 8 show the implied expected values[10] for each respective group from participants' reported joint probability distributions before and after seeing the information. The table displays participants' average prior and posterior for each of the two possible information treatments (lower or higher prevalence) separately, since the rational direction of updating differs depending on the information received. Comparing the average priors to the information provided in each row confirms that in all cases, the lower prevalence information was less than the average prior and the higher prevalence information was greater, meaning that both upward and downward belief revision can be observed across all four health condition and demographic group pairs.

In all cases displayed in Tables 7 and 8, individuals significantly update their beliefs

[10]The expected value was calculated using the midpoint of each interval.

based on the information, and the magnitude of this updating is significantly larger for the group to which the information pertained. The rightmost column of the table, "Difference-in-Differences," demonstrates that this difference is statistically significant at the $\alpha = 0.05$ level in every case. Unlike the results of the experiment reported in Section 1.3, participants in the survey appear not to treat the groups as *entirely* separate, updating their beliefs to some extent about the other group after receiving information specific to a particular group. However, they still appear to believe across all contexts that the information about a group is substantially more relevant to that group than to the other group.

In terms of the accuracy of individuals' beliefs, this information is most harmful in cases where it reinforces an existing error in the prior. For example, participants came into the experiment believing that the prevalence of diabetes among white Americans was significantly higher than the prevalence among Asian Americans, even though the true prevalences in the 2019 NHIS are nearly identical (both round to 9%). If participants subsequently saw a random sample of 50 white Americans with a prevalence of 17%, this gap widened, while participants who saw the alternative sample with a prevalence of 5% end up closer to the correct answer.[11] When individuals' prior had no significant difference between the groups, unrepresentative information in either direction resulted in erroneous differences in the posterior.

The results in Tables 7 and 8 come with two important caveats. The first is that both priors and posteriors are quite flat, suggesting that individuals had a large degree of uncertainty. (For visual depictions of the average priors and posteriors for each question, see Appendix Section A.4.1.) It is likely that unrepresentative information would be easier to dismiss or have less of an effect in an environment where individuals were more confident in their prior. However, despite this, the pattern of belief updating on average was sensi-

[11]Interestingly, in this case, participants' asymmetric updating across groups actually worked in their favor when they were shown a subsample of white Americans with a low prevalence of diabetes, because their posterior ended up both closer to the correct answer because of a general overestimate of diabetes rates in the prior *and* because the erroneous perceived difference between white and Asian Americans in the prior was eliminated in the posterior. While this simply means participants reached the correct conclusion for the wrong reason, it does suggest that unrepresentative group-labeled information may actually be *useful* in some contexts as a brute-force way of correcting erroneous priors.

ble, suggesting that participants understood what to do with the information and how to incorporate it into their answers. The second is that in each question, only about half of the sample reported a different posterior than their prior. While this may suggest they sensibly disregarded the information, it could also be an artifact of the survey design, which pre-filled in the participant's prior when asking them to guess a second time. While this approach has the advantage of reducing participants' cognitive load and ensuring that changes in their answers are conscious and intentional, it also allowed participants who simply wanted to get through as quickly as possible to proceed without thinking. However, if this is the case, it simply biases the differences in Tables 7 and 8 toward zero.

Figure 4 graphically depicts the belief updating patterns among survey respondents whose posterior differed from their prior. In each figure, the dashed line arrow provides a benchmark for respondents' belief updating if they had replaced their prior with exactly the information they were provided (i.e., placed 100% weight on the signal). Across the four health condition and demographic group pairs, in-group updating based on the information (i.e., the difference between the prior and posterior divided by the difference between the prior and the information) varies from 38% to 92%, while out-group updating varies from 4% to 38%.

Table 9 shows that in 3 out of 4 cases, the degree of asymmetric updating between groups was largest for individuals who are themselves members of the group to which the information pertained. Put another way, when shown information about their own group, individuals responded more strongly to the information on average, *and* their response was more skewed toward their own group. This demonstrates an intuitive corollary of the results of the experiment reported in Section 1.3: while in the experiment, individuals treated information about their own group as highly relevant to them and information about other groups as irrelevant, in the survey, individuals treated information about their own group as highly relevant to their group and less relevant to other groups.

The only case in which this did not apply was for college graduates in the question about the prevalence of anxiety. There are two potential intuitive reasons for this. First, when asked at the end of the survey how strong a predictor of health outcomes each of the demographic groups covered in the survey was, educational attainment scored significantly lower than the others.[12] As a result, college graduates responding to the survey might have been less likely to think the information about other college graduates was representative of their own situation, making the information less personally relevant. Second, when compared to high school graduates who did not go to college, college graduates likely have more experience with statistical reasoning and therefore may be more likely to realize the information is noisy and should not strongly influence their beliefs.

While Table 9 provides strong suggestive evidence of an amplifying effect of group membership on asymmetric responses to group-labeled information, there are limitations to keep in mind. Perhaps most importantly, information was only given about one group in each demographic category. As a result, it is not possible to explicitly rule out that the groups selected for the information—women, white Americans, and 18-29 year olds—are simply more likely than other groups to update based on noisy information. Future iterations of the survey which include information about other groups within the same demographic categories can help address this question more conclusively. Additionally, because the pool of survey respondents was disproportionately young, white, and female, it is difficult to conclusively disentangle the effect of membership in a particular group from other overlapping group identities.

1.5.1 Comparing Observed Belief Updating to Bayesian Predictions

In this section, I discuss how survey participants' observed belief updating compares to predicted Bayesian posteriors based on the true variation in the NHIS data among random subsamples of 50 individuals.[13] Unlike the information provided in the experiment described in Section 1.2, this information has nonzero informational content, meaning that a Bayesian

[12]Each demographic group (age, gender, race, and educational attainment) was scored on a scale from 1 to 5, where 1 was the weakest predictor and 5 was the strongest. For educational attainment, the average score was 2.77; for race/ethnicity, 3.32; for age group, 3.77; and for gender, 3.45.

[13]For more details on exactly how the Bayesian posteriors were calculated, see Appendix Section A.2.1.

participant would be expected to update their beliefs after receiving it. In particular, because the condition and demographic group pairs in the survey were deliberately selected to be cases in which expected outcomes are not substantially different across groups, a Bayesian participant with well-calibrated beliefs about group differences would treat the information as equally informative about both groups.[14]

The results in Table 10 indicate that participants' observed belief updating was not consistent with Bayesian predictions based on well-calibrated beliefs about group differences. The highly significant coefficient ($p < 0.01$) on the indicator variable for the group represented in the information strongly rules out the hypothesis that participants on average treated the information as equally informative about both groups. Column 2 adds an interaction term to assess differences in asymmetric belief updating across groups based on whether the participant was a member of the group to which the information pertained. While the significant positive interaction term is consistent with Table 9 in that membership in the group to which the information pertained widens the average asymmetry in belief updating, the non-interaction term remains positive and significant, meaning that participants in other groups still systematically perceived the information as more informative about one group than the other. Columns 3 and 4 of Table 10 repeat the specifications in columns 1 and 2 restricting to only participants who updated their beliefs after the information. Thus, unlike the coefficients in columns 1 and 2 which are pulled toward zero by the inclusion of participants who did not update their beliefs based on the information, the coefficients in columns 3 and 4 represent the magnitude of asymmetric updating and deviations from

[14] Whether or not the *level* of updating is Bayesian is less clear because participants' beliefs about the informativeness or representativeness of the information were not elicited. To carefully avoid both deception and experimenter demand effects, the specific process by which the information was generated was not revealed to participants; they were simply told that it was "a sample of 50 randomly selected [women/white Americans/college graduates/18-29 year olds]." As a result, participants could have had varying beliefs about the representativeness of the information.

Bayesian predictions among updaters specifically.[15] Finally, while the interaction term with the respondent's group membership in column 4 remains positive, it is no longer statistically significant ($p = 0.107$), suggesting that the amplifying effect of group membership is predominantly on the extensive margin of belief updating rather than the intensive margin.

These results comparing observed posteriors to Bayesian predictions are explored in more detail in the Appendix. Appendix Section A.2.1 explains the exact process by which the Bayesian posteriors used in this section were calculated. Appendix Section A.2.2 reports the results of an alternative specification based on a different assumption about individuals' perception of the representativeness of the information. Finally, Appendix Section A.4 provides figures which compare observed posteriors to Bayesian predictions in each condition and demographic group pair, information treatment, and Bayesian posterior specification, respectively.

1.6 Conclusion

In a tightly-controlled lab experiment, I find evidence that individuals have biased beliefs about the relative informativeness of disaggregated information, assuming that information about an individual in a particular category is more informative about others in that category than it is about others in different categories. This bias is apparent even when the underlying data-generating process that assigns individual prospects is explicitly explained to be idiosyncratic. As a result, when the information they receive is noisy, their posterior beliefs frequently feature exaggerated differences between categories, which in turn skews their beliefs about their own prospects. Subsequently acting on these beliefs results in suboptimal choices the individual would not have made if their beliefs had been correct. In the

[15]The results in this table include participants whose priors assigned zero probability to the interval which contained the signal. Many of these participants assigned a nonzero probability to the same interval in their posterior, which is always considered over-inference relative to the Bayesian prediction since Bayes' rule requires the posterior to assign zero probability to any event which had been assigned zero probability in the prior. For an alternative version of the table that excludes individuals with zero priors, see Appendix Table 18.

context of the experiment, this meant that individuals incorrectly guessed their prospects in a lottery with an unobserved, randomly assigned chance of winning, and subsequently took on a level of risk that was not consistent with their general risk preferences.

In a supplemental survey, I apply the insights of the experiment to real-world health-related information. In contexts where there is no true meaningful difference between categories, I find that individuals apply significantly more weight to the information when updating their beliefs about the category reflected in the information than they do when updating their beliefs about other categories. As a result, after observing the information, they frequently end up moving further from the correct answer than their prior, especially when the information reinforces an existing error in their prior. In the context of the survey, this meant that participants frequently ended up thinking two demographic groups had different likelihoods of developing a particular health condition when there was no true significant difference. Even when their beliefs coincidentally move in the right direction, it is for the wrong reason. I also find suggestive evidence of an amplifying effect of personal relevance on the response to information, as individuals appear to update their beliefs especially strongly when the information pertains to their own category. Finally, by simulating Bayesian posteriors using the true sampling variation in the National Health Interview Survey data used to calculate prevalences, I rule out the possibility that the observed belief updating patterns are consistent with a model of Bayesian updating based on well-calibrated beliefs about group differences.

These results demonstrate that individuals may respond to disaggregated information in ways not predicted by models of belief updating and decision-making under uncertainty if those models do not take into account individuals' subjective beliefs about the relative informativeness of each component of a disaggregated signal. In a fully rational model of belief updating which assumes the receiver has correct beliefs about the (lack of) significant heterogeneity in prospects across categories, disaggregated information weakly dominates aggregated information conditional on sample size; in cases where there is significant heterogeneity, disaggregated information provides a less biased estimate of each category's prospects, and in cases where there is not, the rational receiver will simply re-aggregate the information by applying the same weight to each component of the signal. Across the experiment and the survey reported in this paper, participants erroneously assumed that there was significant heterogeneity when there was not, even when they directly observed the individually idiosyncratic process from which prospects were assigned. As a result, in order to accurately predict the observed belief updating in the experiment and survey and subsequent choices in the experiment, one must account for the possibility of a bias in the respective weights receivers apply to each component of the disaggregated signal.

In terms of policy implications, the results of the experiment and survey are a cautionary tale to consider when providing information that informs risky behaviors or the take-up of risk mitigation strategies. Disaggregations of information by age, sex, race, and other first-order demographic characteristics are ubiquitous, and it is likely instinctual to provide them whether or not a difference between subgroups is known or likely. In contexts where group membership strongly predicts outcomes and each subgroup is sufficiently large to limit the influence of noise, such disaggregated information clearly dominates aggregate information for individuals trying to determine their own prospects. However, if either of these conditions are not met, disaggregated information may be worse, as those who receive it may inadvertently become convinced of differences that do not exist. This may be especially true when the information about the receiver's own group is inaccurate based on the results in Table 9 or when the errors in the information are directionally similar to errors in the individual's prior.

While the results in this paper implicate the receivers of the information in multiple overlapping errors in statistical reasoning, it is likely more productive to change the information that is provided than to try to correct receivers' processing of it. The interpretation of disaggregated information is nontrivial: it requires not only an understanding of sample bias and uncertainty as a function of sample size, but also a correct perception of the correlations across categories, especially when information is incomplete or unavailable for particular categories of interest. Given the well-documented difficulties people tend to have with accurately perceiving sample biases and correlations (Ellis and Piccione 2017, Enke 2020, Enke and Zimmermann 2019, Hamill et al. 1980, Kallir and Sonsino 2009, Levy and Razin 2015a,a, Tversky and Kahneman 1974), this is unlikely in general. As a result, providers of information should take care when disaggregating information, only providing it when the average receiver is likely to draw correct conclusions from it.

As a proof-of-concept, I use the experiment data to conduct a back-of-the-envelope counterfactual exercise supposing that group labels had not been provided with the information in the decision stage. For this exercise, I assume that, in the absence of group labels, the effect of observing any additional test run win would be equal to the observed effect of observing an in-group test run win . In practice, I fit an alternative version of the regression in column 1 of Table 3 which replaces the number of in-group wins with the percentage of in-group wins, and then use the resulting coefficients to predict the counterfactual believed X when replacing the percentage of in-group wins with the percentage of overall wins, regardless of group. The result suggests that the average deviation (in absolute value) from the optimal guess of $X = 50$ would have decreased by about 24%.[16] While this is still not the ideal of everyone ignoring the information and simply guessing $X = 50$, participants' beliefs would have been closer to the correct answer on average if they were encouraged to take a sample of 6 random draws into account instead of 3 simply because increasing the sample size reduces the noise.[17]

While the results of the experiment provided suggestive evidence of in-group altruism/favoritism and group identification/attachment as moderators of selective attention to in-group information, further research is needed to pin down the exact mechanisms which explain the observed effect. Further research may also extend the design of the survey described in Section 1.4 to other contexts and group structures to assess generalizability, particularly in

[16]Because the average deviation from the mean is mechanically lower in a prediction from a linear model than in the actual data, this compares the prediction using the original data to the prediction using the counterfactual data to avoid overestimating the reduction.

[17]Note that this is only true because every participant's X was assigned from the same distribution, meaning that increasing the sample size leads to convergence to the true expected value of the distribution from which the individual's X was drawn (i.e., the law of large numbers).

environments where individuals have stronger priors. A greater understanding of how selective attention to in-group information influences individual beliefs across contexts would contribute to multiple areas of interest to social scientists, including systematic differences in behavior and outcomes across groups (e.g., the "white male effect" in social psychology[18]) as well as the formation of stereotypes and discriminatory beliefs.

Future research may also focus on individuals' preferences between aggregated and disaggregated information in contexts where they choose which type of information to receive. In the experiment and survey reported in this paper, participants always received disaggregated information. In this way, the experiment and survey represent contexts in which information is designed or curated in a particular way over which the receiver does not have control. While this is representative of many real-world contexts, such as an individual receiving daily press releases from their county's health department about the spread of COVID-19 or reading a report from their local law enforcement agency about crime in their area, there are also contexts in which the receiver has more direct influence over the structure of the information, either by creating the information themselves from raw data or by choosing between multiple available sources with different structures. The results of this paper suggest that, conditional on receiving disaggregated information in contexts where prospects do not meaningfully differ across categories, individuals process this information in a biased way. If this bias also applies to endogenous selection of information—i.e., individuals believe disaggregated information is better because they expect heterogeneity where there is none and thus believe disaggregated information is more useful than aggregated—this would imply that simply changing the level of aggregation of a single information source would not be sufficient to prevent incorrect beliefs about categorical differences if receivers respond to this change by seeking out disaggregated alternatives. In this case, direct interventions to debias individuals' beliefs about heterogeneity and the relative informativeness of each component of a disaggregated signal may be necessary.

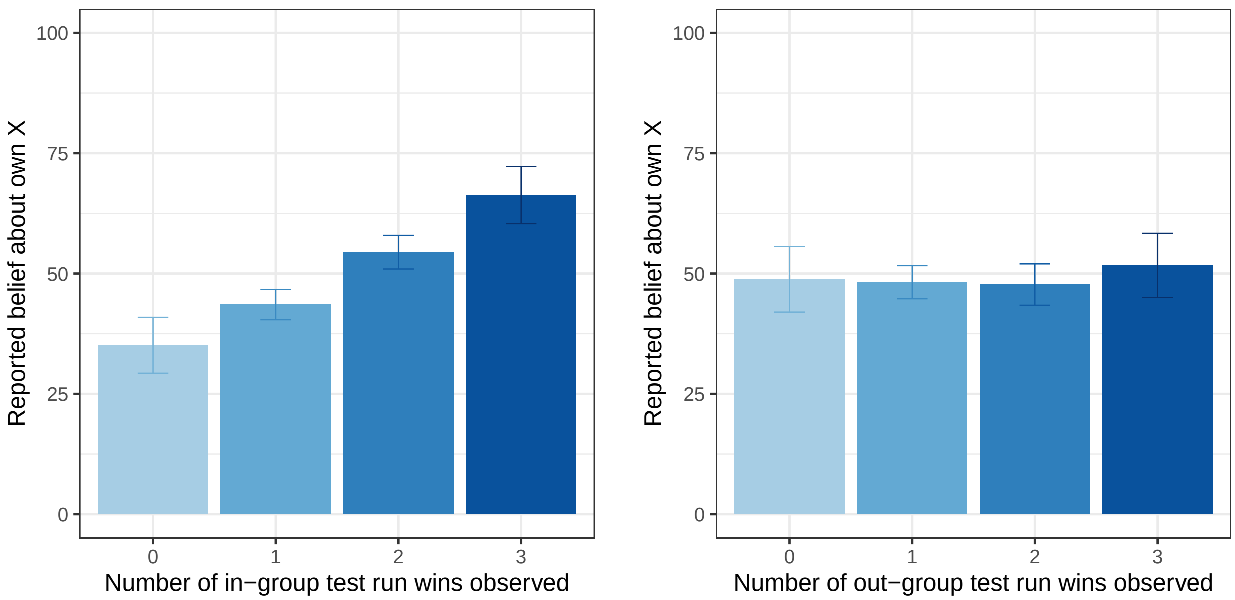
[18] e.g. Finucane et al. (2000), Flynn et al. (1994), Palmer (2003).

1.7 Figures

Figure 1: Relationship Between Beliefs About X and Test Run Wins Observed

(a) Response to in-group information

(b) Response to out-group information



Notes: "Number of in-group (out-group) test run wins observed" refers to the three in-group (out-group) hypothetical draws of the lottery with X chance of winning \$5 (based on others' unobserved assigned X values) which participants observed prior to guessing their own X. Participants were informed that these draws were for informational purposes only and did not affect others' payoffs. "Reported belief about own X" was reported on a slider over integers between 0 to 100.

Figure 2: Session-Level Beliefs about Men and Women by Test Run Wins Observed

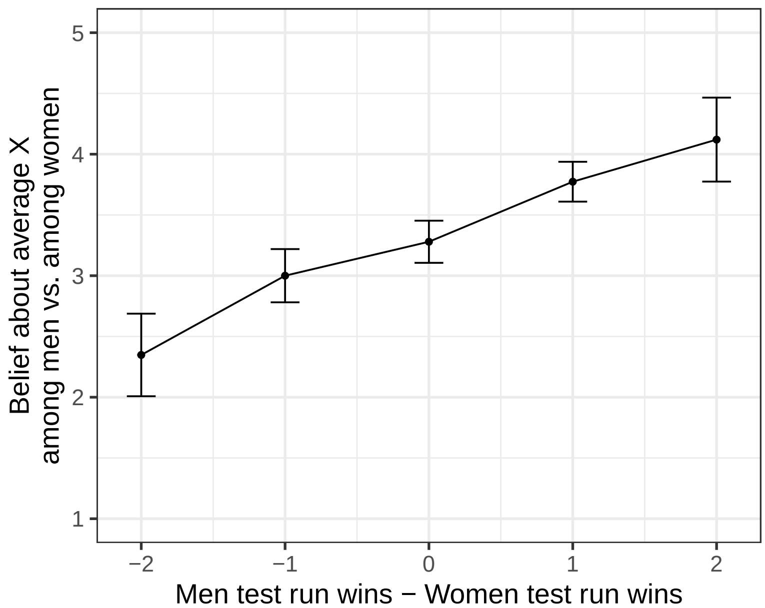
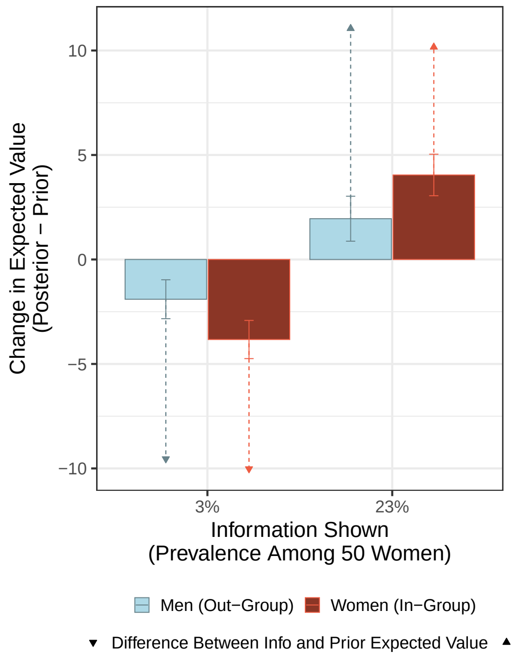


Figure 3: Example of a Joint Probability Distribution Matrix from the Survey

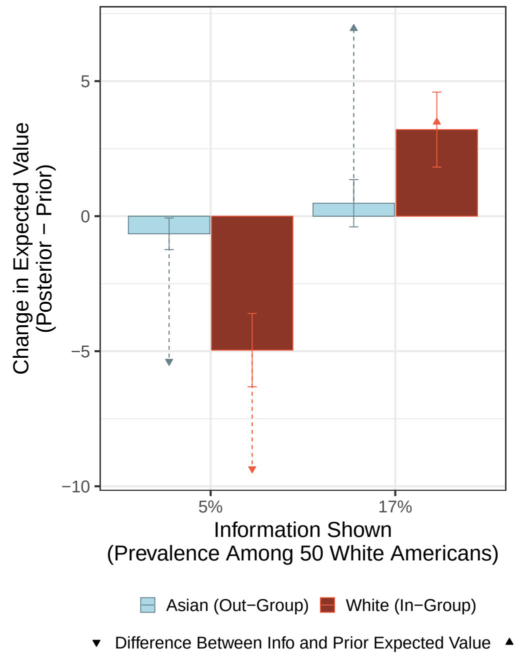
		Women		
		Between 0 and 8% have had cancer	**Between 9 and 16%** have had cancer	**Between 17 and 25%** have had cancer
Men	**Between 0 and 8%** have had cancer	A	D	G
	Between 9 and 16% have had cancer	B	E	H
	Between 17 and 25% have had cancer	C	F	I

Figure 4: Survey Belief Updating Figures (Updaters Only)

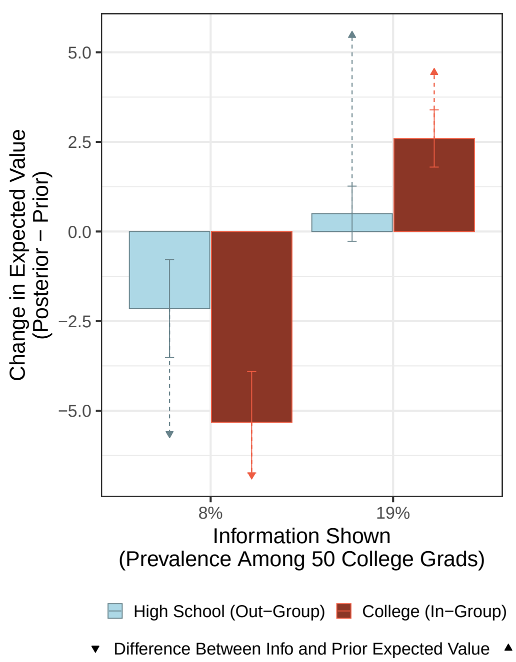
(a) Cancer by Gender



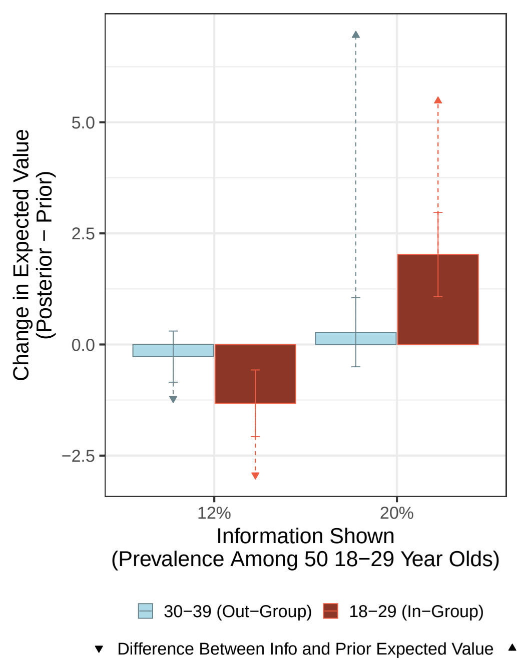
(b) Diabetes by Race



(c) Anxiety by Education



(d) Depression by Age Group



1.8 Tables

Table 1: Example of Decision Stage Test Run Information

Your group of men's test run results:	**Other group of women's test run results:**
Man 1: Win	Woman 1: Lose
Man 2: Lose	Woman 2: Win
Man 3: Lose	Woman 3: Win

Table 2: Decision Stage Price List

Option A	**Option B**
X% chance of \$5, (100-X)% chance of \$0	\$0.50 for sure
X% chance of \$5, (100-X)% chance of \$0	\$1.00 for sure
X% chance of \$5, (100-X)% chance of \$0	\$1.50 for sure
X% chance of \$5, (100-X)% chance of \$0	\$2.00 for sure
X% chance of \$5, (100-X)% chance of \$0	\$2.50 for sure
X% chance of \$5, (100-X)% chance of \$0	\$3.00 for sure
X% chance of \$5, (100-X)% chance of \$0	\$3.50 for sure
X% chance of \$5, (100-X)% chance of \$0	\$4.00 for sure
X% chance of \$5, (100-X)% chance of \$0	\$4.50 for sure
X% chance of \$5, (100-X)% chance of \$0	\$5.00 for sure

Table 3: Regressing Beliefs and Switch Point Choices on Test Run Wins

	(1) Believed Own X	(2) Switch Point	(3) Believed Own X	(4) Switch Point
Observed in-group test run wins	10.435***	0.441***	10.899***	0.546**
	(1.213)	(0.151)	(1.850)	(0.231)
Observed out-group test run wins	0.189	0.095	1.605	-0.007
	(1.188)	(0.148)	(1.728)	(0.216)
In-group wins × In-group information on left			-0.866	-0.171
			(2.504)	(0.313)
Out-group wins × In-group information on left			-2.659	0.199
			(2.420)	(0.302)
Male indicator	5.077**	0.417	4.678**	0.430
	(2.121)	(0.264)	(2.148)	(0.268)
Age in years	-0.221	-0.012	-0.311	-0.021
	(0.424)	(0.053)	(0.440)	(0.055)
Risk tolerance	-0.045	0.145*	-0.102	0.143*
	(0.661)	(0.082)	(0.667)	(0.083)
Observations	232	232	232	232
Adjusted R^2	0.245	0.047	0.241	0.038

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Group-clustered standard errors in parentheses. "Number of in-group (out-group) test run wins observed" refers to the three in-group (out-group) hypothetical draws of the lottery with X chance of winning $5 (based on others' unobserved assigned X values) which participants observed prior to guessing their own X. Participants were informed that these draws were for informational purposes only and did not affect others' payoffs. "Reported belief about own X" was reported on a slider over integers between 0 to 100. "In-group information on left" is an indicator variable which equals 1 when the participant saw the information about their own group in the left-hand column of the table in which test run results were reported. For an example of this table, see Figure 1. All regressions include controls for priming conditions explained in Appendix Section A.1.

Table 4: Group Identification as a Moderator of Response to In-Group Information

	Believed X	Believed X
Observed in-group test run wins	4.226	7.382***
	(2.904)	(2.357)
In-group wins × Amount allocated to in-group receiver	9.761**	
	(4.278)	
In-group wins × Group identification		0.804*
		(0.486)
Observed out-group test run wins	0.350	0.312
	(2.859)	(1.840)
Out-group wins × Amount allocated to in-group receiver	-0.367	
	(4.315)	
Out-group wins × Group identification		0.038
		(0.365)
Amount allocated to in-group receiver	-8.556	
	(11.093)	
Self-reported group identification		-1.763
		(1.123)
Observations	232	232
Adjusted R^2	0.256	0.253

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Group-clustered standard errors in parentheses. "Num-ber of in-group (out-group) test run wins observed" refers to the three in-group (out-group) hypothetical draws of the lottery with X chance of winning \$5 (based on others' unobserved assigned X values) which participants observed prior to guessing their own X. Participants were informed that these draws were for informational purposes only and did not affect others' payoffs. "Reported belief about own X" was reported on a slider over integers between 0 to 100. "Amount allocated to in-group receiver" refers to the amount participants allocated to the in-group receiver when asked to allocate \$1 between an in-group receiver and an out-group receiver. "Group identification" refers to the participants' response of how attached they felt to their group throughout the experiment on a scale from 1 to 10. Controls for gender, age, and priming condition are included.

Table 5: Session-Level Belief Regressions

Dependent variable: Belief about relative average X for men and women *(5-point Likert scale, 1 = Women's X was much higher than men's; 5 = Men's X was much higher than women's)*		
	(1)	(2)
Men test run wins - Women test run wins	0.402***	0.503***
	(0.041)	(0.058)
Men test run wins - Women test run wins $\times$ Male indicator		-0.196**
		(0.079)
Male indicator	0.032	0.032
	(0.103)	(0.099)
Age in years	0.021	0.021
	(0.017)	(0.016)
Observations	231	231
Adjusted R^2	0.288	0.302

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Group-clustered standard errors in parentheses. "Men test run wins - Women test run wins" refers to the difference in outcomes (an integer between -3 and 3) between men and women among the six hypothetical draws of the lottery with X chance of winning $5 (based on others' unobserved assigned X values) which participants observed prior to guessing their own X. For example, if the participant observed 2 wins among men and 1 win among women, this variable takes a value of 2 - 1 = 1. Participants were informed that these draws were for informational purposes only and did not affect others' payoffs. The "Belief about average X for men and women" was reported on a 5-point Likert scale from "Women's X was much higher than men's" to "Men's X was much higher than women's" and then assigned an integer value between 1 and 5. Note that these regressions have one fewer observation (231 vs 232) than previous tables because one participant disconnected from the session before answering this question. All regressions included controls for the priming condition. For explanations of the priming conditions, see Appendix Section A.1.

Table 6: Survey Sample Demographics

Gender		
	Male	0.40
	Female	**0.58**
	Non-binary or third gender	0.02
Race/Ethnicity		
	Arab/Middle Eastern	0.00
	Asian/Pacific Islander	0.06
	Black/African American	0.12
	Caucasian/White	**0.61**
	Hispanic/Latino	0.07
	Multiple races or other	0.13
Age Group		
	18-29	**0.58**
	30-39	0.28
	40+	0.13
Education		
	Less than high school	0.01
	High school graduate	0.07
	Some college	0.26
	Undergraduate degree (associate's or bachelor's)	**0.36**
	Advanced degree (Master's or higher)	0.29

Table 7: Prior and Posterior Expected Values by Group: Cancer and Diabetes

Cancer (Correct Answers: Men 9%, Women 10%)

Low Prevalence Information (Women = 3%)

Prior		Posterior		Difference-in-Differences
Men	Women	Men	Women	
12.69	12.91	11.51	10.58	-1.15**
(3.13)	(2.88)	(3.10)	(3.58)	

High Prevalence Information (Women = 23%)

Prior		Posterior		Difference-in-Difference
Men	Women	Men	Women	
12.24	13.36	13.27	15.48	1.09**
(3.13)	(2.88)	(3.10)	(3.58)	

Anxiety (Correct Answers: White Americans 9%, Asian Americans 9%)

Low Prevalence Information (White = 5%)

Prior		Posterior		Difference-in-Differences
White	Asian	White	Asian	
13.88	11.08	10.97	10.70	-2.52***
(3.13)	(3.36)	(4.10)	(3.34)	

High Prevalence Information (White = 17%)

Prior		Posterior		Difference-in-Difference
White	Asian	White	Asian	
13.88	11.11	15.22	11.23	1.22**
(3.13)	(2.88)	(3.10)	(3.58)	

Table 8: Prior and Posterior Expected Values by Group: Anxiety and Depression

Anxiety (Correct Answers: High School Grads 14%, College Grads 14%)

Low Prevalence Information (College Grads = 8%)

Prior		Posterior		Difference-in-Difference
High School Grads	College Grads	High School Grads	College Grads	
13.31	14.30	12.20	11.56	-1.64***
(3.19)	(2.70)	(2.95)	(3.89)	

High Prevalence Information (College Grads = 19%)

Prior		Posterior		Difference-in-Differences
High School Grads	College Grads	High School Grads	College Grads	
13.39	14.09	13.63	15.40	1.07***
(3.34)	(2.65)	(2.98)	(3.58)	

Depression (Correct Answers: 18-29 Year Olds 16%, 30-39 Year Olds 15%)

Low Prevalence Information (18-29 Year Olds = 12%)

Prior		Posterior		Difference-in-Differences
18-29 Year Olds	30-39 Year Olds	18-29 Year Olds	30-39 Year Olds	
14.08	13.32	13.45	13.16	-0.48**
(2.74)	(2.44)	(2.10)	(2.52)	

High Prevalence Information (18-29 Year Olds = 20%)

Prior		Posterior		Difference-in-Difference
18-29 Year Olds	30-39 Year Olds	18-29 Year Olds	30-39 Year Olds	
13.96	13.17	14.97	13.33	0.85**
(3.40)	(3.00)	(4.01)	(3.00)	

Table 9: Effect of Membership in Group to Which Information Pertained on Asymmetric Belief Updating Patterns

		Average Diff-in-Diffs[19]		
Condition	Group Represented in Information	In-Group	Out-Group	p-value
Cancer	Women	1.86	0.89	0.023**
Diabetes	White[20]	2.25	1.25	0.086*
Anxiety	College Grads	1.21	1.42	0.673
Depression	18-29 Year Olds	0.98	0.23	0.026**

[8]Before taking the average, the difference-in-differences was multiplied by -1 in the case of low prevalence information.

[9]Because races are not mutually exclusive, the definition of "in-group" in this case is ambiguous. For this comparison, only individuals who exclusively identified as white are considered "in-group." (Survey participants were able to select multiple options in the question about race/ethnicity.) This is in keeping with the NHIS from which the correct answers were drawn, which classifies individuals of multiple races as separate groups.

Notes: "Difference-in-Differences" reports the following statistic: (Average Posterior Expected Value for Group Represented in Information - Average Prior Expected Value for Group Represented in Informa-tion) - (Average Posterior Expected Value for Other Group - Average Prior Expected Value for Other Group). The stars indicate the p-value comparing this difference to 0 (* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$). This table includes all 202 participants who responded to the survey in all reported statistics.

Table 10: Comparing Observed Posteriors to Bayesian Predictions: Baseline Model

	Dependent variable: Difference (error) in probability assigned to interval containing the information between observed posterior and Bayesian prediction			
Constant	-0.014	-0.019	0.030*	0.037
	(0.010)	(0.014)	(0.016)	(0.026)
Group in info	0.124***	0.086***	0.227***	0.197***
	(0.014)	(0.020)	(0.022)	(0.036)
Participant in group in info		0.009		-0.012
		(0.020)		(0.032)
Group in info X Participant in group in info		0.071**		0.047
		(0.028)		(0.046)
Sample	All	All	Updaters	Updaters
Num.Obs	1616	1616	834	834
R2 Adj.	0.046	0.054	0.111	0.111

Notes: * p < 0.1, ** p < 0.05, *** p < 0.01. The constant represents the average error relative to the Bayesian predicted posterior in the group not represented in the information. Negative values indicate under-inference while positive values indicate over-inference. "Group in info" is an indicator variable which equals 1 when the observation contains a belief about the group to which the information pertained. For example, in the cancer prevalence by gender questiion, "Group in info" equals 1 for beliefs about women and 0 for beliefs about men. "Participant in group in info" is an indicator variable which equals 1 when the *participant* to which the observation pertains is in the group to which the information pertained, regardless of which group the belief is about. In the same cancer by gender example, "Participant in group in info" equals 1 for all female participants' beliefs about men *and* women. For an explanation of how these Bayesian predictions were calculated, see Appendix Section A.2.1.

2.0 Water Availability and Heat-Related Mortality: Evidence from South Africa

Rising global surface temperatures threaten to reduce precipitation and evaporate surface freshwater in areas already experiencing water stress. In this paper, I demonstrate that higher upstream water availability significantly reduces the slope of the temperature-mortality relationship during the summer. This suggests investment in water infrastructure is an effective community-level adaptation to climate change, especially where the status quo of water access is relatively poor. As an example of such investment, I show a transnational water transfer project both increased water availability and reduced hot-season mortality in receiving districts.

2.1 Introduction

As global surface temperatures rise, adaptations to heat become more important to survival and quality of life. Excess heat has been shown to decrease cognitive performance (Zivin et al. 2018), increase cardiovascular and respiratory mortality risk (Basagaña et al. 2011, Curriero et al. 2002), increase incidence and severity of injury during physical exertion (Nelson et al. 2011), increase incidence of low birth weight (Deschênes et al. 2009) and infant mortality (Banerjee and Bhowmick 2016), and ultimately, increase overall mortality (Hajat and Kosatky 2010). Higher temperatures have also been associated with reduced economic production through effects on time use (Graff Zivin and Neidell 2014), crop yields (Schlenker and Roberts 2009), and aggregate economic activity (Burke and Emerick 2016).

In this paper, I demonstrate that higher potable water availability significantly reduces the slope of the heat-mortality relationship. At the status quo median of water availability, I find a statistically significant, positive relationship between heat and mortality, with effect sizes in line with prior literature (Burgess et al. 2017, Hajat et al. 2005). However, at one standard deviation above the mean of water availability, the heat-mortality relationship is not statistically significant, with a substantially smaller point estimate. For example, a back-of-the-envelope calculation suggests the December 2018 heat wave in Pretoria[1], despite only lasting a few days, would on average increase the monthly mortality rate by about 3.2 per million people ($p < 0.01$). Increasing potable water availability by one standard deviation from the mean reduces this point estimate to about 0.8 per million ($p \approx 0.29$).

I employ two causal identification strategies. First, I create measures of upstream, downstream, and within-district potable water availability for each of the 52 districts of South Africa, and I isolate the effect of upstream water availability by controlling for within-district and downstream measures in panel fixed-effects regressions. This strategy has been used in prior literature (e.g. Jerch (2018), Chakraborti (2016), Garg et al. (2018)) to address confounders such as local precipitation, which affects a broad range of local outcomes that may be correlated with the heat-mortality relationship (e.g. areal flooding, land suitability for agriculture). To confirm that the estimated effect is specific to *heat*-related mortality, I use colder months as a comparison group in a difference-in-difference specification, finding that upstream water availability differentially reduces mortality in the summer. Finally, I introduce local precipitation controls and estimate conditional moderating effects of water availability at varying levels of precipitation. I only find a significant moderating effect when local precipitation is relatively scarce, suggesting that upstream water availability moderates heat-related mortality by insuring against local drought.

Second, in the Online Appendix, I use a transnational water transfer project as a natural experiment that increased potable water availability in receiving districts. The Lesotho Highlands Water Project, formally inaugurated in 2004, diverted water from the mountains of Lesotho to Gauteng Province, the densely populated industrial center of South Africa. In doing so, the transfer created a new way for upstream water sources to reach targeted

[1] Coverage: https://www.thesouthafrican.com/weathersa/gauteng-weather-heatwave-expected/. The heat wave resulted in 5 cooling degree days (CDD) in Pretoria in December 2018 at a base temperature of 90°F (used throughout this paper as a measure of heat incidence).

districts as well as those positioned downstream. In a difference-in-difference specification, I find the slope of the heat-mortality relationship differentially declined in receiving districts after 2004, and this difference cannot be explained by a decline in overall mortality. In addition to corroborating the preceding findings of the paper, this suggests investment in water infrastructure is an effective community-level adaptation to climate change, especially where the status quo of water access is relatively poor.

This finding contributes to the growing literature on the efficacy of adaptations. Adaptations to avoid heat damages can be placed in two broad categories: household-level and community-level (Deschenes 2014). Household-level adaptations are predominantly based on heat avoidance, including spending more time indoors; investing in fans, better insulation, or air conditioning; and migrating away from the heat. Households engaged in agricultural production can also adapt through crop choice and irrigation (Burke and Emerick 2016, Di Falco et al. 2011). Community-level adaptations include early-warning systems for extreme weather, building climate-controlled shelters, and increased access to quality water (Deschenes 2014). This paper provides evidence that increased potable water availability, which has already been shown to have several other positive effects (e.g. Ao (2016), Devoto et al. (2012)), is an effective adaptation to heat.

The paper proceeds as follows. Section 2 describes the data and context. Section 3 describes the empirical strategy used to identify the causal effect of water availability on the heat-mortality relationship. Section 4 presents results. Section 5 uses a water transfer project as a natural experiment increasing water availability in receiving districts. Section 6 concludes.

2.2 Data

To identify the effect of water availability on the heat-mortality relationship, I have constructed a panel of mortality, temperature, hydrological, and geographic data. I describe each component of this panel below.

Mortality data. I obtained counts of deaths by district and month from 1997 to 2015 from Statistics South Africa. Since there are many ways excess heat can increase mortality, including unnatural causes (see Dell et al. (2014) for a review), my dependent variable includes all deaths, regardless of cause. For a robustness check, I also obtained counts of deaths attributed to infectious gastroenteritis and diarrhea, a leading waterborne cause of death in sub-Saharan Africa, to verify that the results hold for causes of mortality more directly associated with lack of quality water.

Temperature data. I use cooling degree days (CDD) at a base temperature of 90°F (with other base temperatures as robustness checks) as a measure of heat. Degree days are primarily a measure of the energy required to cool or heat a building's interior to the base temperature[2], but prior literature in environmental economics has used them as a measure of heat exposure (e.g. Deschênes and Greenstone (2007)). I use CDD in this paper for two reasons. One, it is a continuous monthly measure of both duration and intensity of heat, which may have independent or cumulative effects on mortality. Two, as depicted in Figure 5, the marginal effect of heat on the mortality rate is increasing as the temperature rises above 70°F, and there is a wide range of temperatures for which the marginal effect is zero or negative. Because CDD truncates temperatures below the selected threshold, the estimated coefficient regressing CDD on mortality rates is the average marginal effect of temperature above the threshold (i.e., the effect of excess heat), which is the effect of interest to this paper.

I construct monthly CDD measures from the Global Historical Climatology Network Daily (GHCN-Daily) dataset provided by the National Oceanic and Atmospheric Association (NOAA). The spatial distribution of weather stations included in this dataset are described in Section 2.3.1. The base temperature was selected in line with prior literature, which typically defines "excess heat" as temperatures above the 90th or 95th percentile in a region (Burgess et al. 2017, Curriero et al. 2002, Hajat and Kosatky 2010). To ensure that the

[2]Mathematically, one CDD 90°F is equivalent to one day (24 hours) in a month during which the average temperature exceeded 90°F by one degree.

findings are robust to more flexible measures of heat exposure, I also calculate CDD at a base temperature of 75°F. I combine the geographic coordinates of each weather station with the GIS data described below to create a spatiotemporal measure of heat exposure.

Hydrological data. The South African Department of Water and Sanitation (DWS) maintains an online portal of hydrological data collected from dams and monitoring stations. I retrieved all available water storage levels from 1996 to 2016, resulting in a sample of 773 dams and reservoirs. This data includes the geographic coordinates of each dam, which I combine with GIS data described in the next paragraph to construct a spatiotemporal measure of upstream, within-district, and downstream water availability by district. A map of all dams in the sample and their associated catchment areas is provided in Figure 2.

GIS data. A digital elevation model (DEM) for South Africa was obtained from the Consortium for Spatial Information (CGIAR-CSI), based on raster data from the NASA Shuttle Radar Topography Mission (SRTM). Elevation is recorded at a 3 arc-second (90 meter) resolution. A shapefile of rivers and dams was provided by the Department of Water and Sanitation, and a shapefile of South African district boundaries in 2016 was obtained from the Humanitarian Data Exchange (HDX).

Table 1 presents pertinent summary statistics by province. The high degree of economic inequality in South Africa is mirrored in the disparate life expectancies[3] across provinces and sexes, ranging from just 53 years for men in Free State to 70 years for women in Western Cape between 2011 and 2016 (Stats SA 2018). However, substantial variation in mortality risk across provinces and population subgroups persists. There are also large differences across provinces in average educational attainment, population group composition, elevation, and access to piped water. More than two decades after the end of apartheid, inequality persists along racial lines, with Black Africans in 2015 being 15 times as likely to be HIV+, twice as likely to have no schooling, and less than half as likely to have piped water in-residence

[3]One major driver of life expectancy is the high prevalence of HIV/AIDS, which hit an all-time high of 13.06% in 2018 (22.32% among women age 15-49) according to Stats SA estimates. The increase in life expectancy from 2001-2006 to 2011-2016 is also attributable to the high HIV/AIDS prevalence, since the South African government began ramping up provision of free antiretroviral therapy (ART) drugs for HIV-positive individuals in the mid-2000s (Haal et al. 2018).

compared to their White South African counterparts, according to Statistics South Africa's General Household Survey.

2.3 Empirical Strategy

The spatial distribution of the 68 weather stations included in the Global Historical Climatology Network Daily (GHCN-Daily) dataset with temperature data between 1997 and 2016 is depicted in Figure 6. While there is some clustering of weather stations around population centers and along the southern coastline, there is at least one station within 150 kilometers of the geographic center of each of the 52 districts. To limit the influence of station-specific measurement error, I remove all observations that fail at least one of the internal consistency checks included in the GHCN-Daily data before constructing the distance-weighted average described in Section 2.3.1.

Figure 7 depicts the spatial distribution of the 773 dams, reservoirs, and river flow stations monitored by the Department of Water and Sanitation (DWS) between 1997 and 2016, their catchment areas, and the surrounding area elevation gradient. The hydrogeology and climate of South Africa is largely determined by elevation differences and proximity to the coastline. Smaller dams are scattered across the coastline, where the rivers are densest. The coastal districts are separated from the inland districts by a skewed U-shape escarpment, particularly in the southeast, where the elevation peaks in the sovereign kingdom of Lesotho. The belt of larger catchment area dams through the northwest and center of the country lie along the Orange River and its largest tributary, the Vaal River. The importance of these particular rivers will be explored in more detail in Section 5.

2.3.1 Measure of Heat Exposure

To obtain a measure of heat exposure for each district j at month t, I construct a distance-weighted average of cooling degree days (CDD) at base temperatures of 90°F and

75°F, respectively, from the values observed at each weather station s. The formula for this weighted average is shown in equation 1, where $d(\cdot)$ represents the geodesic distance function[4] (Karney 2013). The distance from each district to each weather station is calculated from the geographic center of the district.

$$CDD_{jt} = \sum_{s} \left(\frac{[d(j,s)]^{-4} CDD_{st}}{\sum_{s} [d(j,s)]^{-4}} \right) \tag{1}$$

I employ a distance-weighted average measure of heat exposure because the death counts I use to calculate mortality rates only identify the location of the decedent at the district level. Thus the specific location within the district of each decedent at their time of death is unobserved. The weights in equation 1 are estimates of the probability that an individual who died in district j during month t was exposed to the average temperature recorded at station s during month t, respectively. This relies on the assumption that the unobserved true temperature experienced by each decedent is a convex combination of contemporaneous observations from the 68 weather stations in Figure 6. If this assumption is violated, or if the distance from the center of a district is not an accurate estimate of the probability an individual in that district experiences the conditions recorded by that weather station (e.g., if population density is concentrated away from the center), this weighted average will introduce measurement error. However, the cooling degree day (CDD) measure of heat exposure is only nonzero if the average temperature exceeds 90°F, which is uncommon in most of South Africa; thus the weighted average is more likely to underestimate heat exposure than to overestimate it, and systematic overestimates caused by using weather data from hotter-on-average areas will be absorbed by fixed effects. If the true heat-mortality relationship is significant and positive above the threshold used to calculate CDD, which has been a robust finding in prior literature, measurement error will attenuate the estimated effect by undercounting the number of deaths that occur during heat waves. Thus this measurement error is likely to bias the heat-mortality relationship coefficient toward zero, and thus bias the

[4]The geodesic distance is the shortest distance between two points on an ellipsoid.

coefficient on interactions between the heat-mortality relationship and the water availability measures described in Section 2.3.2 toward zero as well.

2.3.2 Measure of Water Availability

For each district, I classify each dam as *upstream*, *downstream*, or *within-district*, based on the dam's elevation relative to the distance from the center of the district. This classification allows me to isolate the effect of upstream water availability, which is advantageous to both causal identification and interpretation. In terms of identification, controlling for within-district and downstream water levels absorbs the effect of local precipitation and other local hydrological conditions. Prior studies of water-related shocks and their effect on health outcomes have employed a similar upstream-downstream strategy, such as Jerch (2018) using variation in downstream population as an instrument to identify the effect of Clean Water Act compliance on water quality and resident costs, Chakraborti (2016) using downstream water quality to estimate manufacturing plants' responses to ambient environmental factors, and Garg et al. (2018) using upstream polluting behavior to estimate the effect of water pollution on nearby residents' health.

While local precipitation certainly affects potable water availability, it is simultaneously related to other local covariates such as indoor/outdoor time use, cloud cover (and thus exposure to and intensity of sunlight), and relative humidity, which in turn affect the heat-mortality relationship. Holding within-district and downstream water levels fixed, an increase in upstream water levels increases potable water availability in a district through the natural flow of water downstream without changing these confounders. For this reason, the effect of upstream water availability on the heat-mortality relationship is more plausibly representative of the effect of increases in water supply through infrastructure, which also involves moving water into a district without the other effects of precipitation. Since precipitation is expected to become less frequent and less predictable in South Africa as climate change progresses (Nkhonjera 2017), it is important to confirm the effect of water availability on the heat-mortality relationship is not conditional on the water being delivered through local precipitation.

Let $E(k)$ represent the elevation of dam k and $E(j)$ represent the elevation of the center of district j. Let δ_{ujk}, δ_{wjk}, and δ_{djk} be indicator variables defined by the following equations 2a-2c:

$$\delta_{ujk} = \begin{cases} 1 & \text{if } d(j,k) > 1 \text{ and } E(k) > E(j) + 100; \\ 0 & \text{otherwise} \end{cases} \tag{2a}$$

$$\delta_{djk} = \begin{cases} 1 & \text{if } d(j,k) > 1 \text{ and } E(k) < E(j) - 100; \\ 0 & \text{otherwise} \end{cases} \tag{2b}$$

$$\delta_{wjk} = \begin{cases} 1 & \text{if } d(j,k) < 1 \\ 0 & \text{otherwise} \end{cases} \tag{2c}$$

Let v_{kt} denote the mean recorded level of dam k in month t in standard deviations from its sample mean.[5] I construct V_{ijt}, a distance-weighted average of dam levels in each respective category for each district j, according to equation 3.

$$V_{ijt} = \sum_k \left(\frac{\delta_{ijk} \cdot [d(j,k)]^{-4} \cdot v_{kt}}{\sum_k [d(j,k)]^{-4}} \right) \quad \text{for} \quad i \in \{u, d, w\} \tag{3}$$

Figure 8 provides a graphical representation of this process for the City of Johannesburg, depicting dams by category that were assigned a weight of at least 0.01 in equation 3, with the size of each circle representing the weight. Since Johannesburg is at a high elevation relative to most of the country, the selected downstream dams form a ring around the district,

[5]Dam levels are standardized before averaging because heterogeneity in dams, based on their design capacity, location, and maintenance is endogenous and confounds the effect of water availability.

and the only selected upstream dams are those constructed for the Lesotho Highlands Water Project.

Any effect of dam levels on the heat-mortality relationship in month t is unlikely to be driven by dam levels in month t alone. By design, dams stabilize water supply over time, storing water during positive shocks to river flow volume to insure against future negative shocks. This is especially significant in semi-arid South Africa, which relies on recharge during the rainy season (October to March) to meet water demand throughout the dry season. This seasonal pattern is illustrated by district in appendix figure 41. It also takes time for stored water to pass through sanitation facilities and utility pipelines to end users. Thus I construct a twelve-month lagged median of dam levels, $\overline{V_{ijt}}$, described in equation 4. I use this measure of water availability because it eliminates seasonal variation, which is correlated with seasonal covariates other than heat that may affect mortality (e.g. short-run agricultural income), deemphasizes outliers produced by one-off shocks to dam levels that may also affect mortality (e.g. flooding), and limits the influence of non-systematic measurement error.

$$\overline{V_{ijt}} = \text{med}\left(\{V_{ijt-1},\ V_{ijt-2},\ \ldots, V_{ijt-13}\}\right) \text{ for } i \in \{u, d, w\} \tag{4}$$

2.3.3 Specifications

Let M_{jt} represent the mortality rate in district j in month t and let y represent the fixed effect for the year of month t. Equation 5 describes the regression model used to estimate the effect of potable water availability on the heat-mortality relationship.

$$\begin{aligned} M_{jt} = \beta_0 + \beta_{CDD} CDD_{jt} + \beta_u \overline{V_{ujt}} + \beta_d \overline{V_{djt}} + \beta_w \overline{V_{wjt}} + \gamma_u \left(CDD_{jt} \times \overline{V_{ujt}}\right) + \\ \gamma_d \left(CDD_{jt} \times \overline{V_{djt}}\right) + \gamma_w \left(CDD_{jt} \times \overline{V_{wjt}}\right) + \Omega\left(j \times y\right) + \epsilon_{jt} \end{aligned} \tag{5}$$

The coefficient of interest is γ_u, the moderating effect of upstream water availability ($\overline{V_{ujt}}$) on the heat-mortality relationship coefficient β_{CDD}. I include the downstream ($\overline{V_{djt}}$)

and within-district ($\overline{V_{wjt}}$) dam levels and their interactions with CDD_{jt} in the model as independent regressors so that γ_u isolates the effect of upstream water availability.[6] While it is reasonable to expect a mitigating effect of within-district or downstream water availability as well, γ_d and in particular γ_w are confounded by other local factors related to water availability that are likely to influence the mortality rate, including areal flooding, land suitability for agriculture, and humidity. Holding within-district dam levels fixed, an increase in upstream dam levels at a sufficient distance from the district center increases potable water supply and is unlikely to be related to these confounding factors.

The other controls in equation 5, denoted by $\Omega\left(j \times y\right)$, are district-by-year fixed effects (FE). I include these in all specifications to absorb the effects of unobserved determinants of the mortality outcome variable M_{jt} that changed during the sample period. One such unobserved source of variation is the rollout of free antiretroviral therapy (ART) drugs for HIV-positive individuals in the mid-2000s (Haal et al. 2018). The effect of this rollout is likely to be heterogeneous over time, as the program was rolled out and individuals started taking the medication, and by location, since the proportion of the population living with HIV differs by district. It may also vary by location over time if the rollout differed in timing or take-up across districts. District-by-year FE flexibly absorb the effect of free ART drugs on mortality rates in the presence of any or all of these potential sources of heterogeneity. Since M_{jt} includes all deaths of any cause, this is just one example of many potential unobserved confounders that are absorbed by district-by-year FE. Because district-by-year FE are included in all specifications, the estimated effect of water availability on the heat-mortality relationship is identified by within-district-year variation in temperature, water availability, and mortality rates.

To isolate the relationship between the mortality rate and excess heat, I restrict the sample to the October to March summer to estimate equation 5. The temperature-mortality relationship is typically U-shaped, associating both extreme heat and extreme cold with higher mortality rates. In the years covered by my data, I find a substantially higher average

[6]Upstream, downstream, and within-district dam levels as defined in this paper are correlated, but not perfectly (coefficients ranging from 0.4 to 0.6), as expected.

mortality rate during the cold season than during the summer. Thus including the cold season in the estimation sample for equation 5 results in a negative estimate of β_{CDD}, even at base temperatures as high as 100°F.[7] As a robustness check, I estimate the difference-in-difference model described in equation 6 to confirm the effect of upstream water availability is more significant in the summer (H) using the cold season (C) as a comparison group.

$$M_{jt} = \beta_0' + \beta_H \mathbf{1}_{t\in\mathbf{H}} + \beta_u' \overline{V_{ujt}} + \beta_d' \overline{V_{djt}} + \beta_w' \overline{V_{wjt}} + \gamma_u' \left(\mathbf{1}_{t\in\mathbf{H}} \times \overline{V_{ujt}}\right) + \gamma_d' \left(\mathbf{1}_{t\in\mathbf{H}} \times \overline{V_{djt}}\right) + \gamma_w' \left(\mathbf{1}_{t\in\mathbf{H}} \times \overline{V_{wjt}}\right) + \Omega\left(j \times y\right) + \epsilon_{jt} \tag{6}$$

In this case, the coefficient of interest is γ_u', which will be negative and significant if upstream water availability differentially mitigates the heat-mortality relationship rather than simply decreasing overall mortality. Aside from replacing CDD_{jt} with an indicator variable that equals 1 when month t is between October and March ($\mathbf{1}_{t\in\mathbf{H}}$), equation 6 is identical to equation 5.

2.4 Results

Figures 9 and 10 graphically represent the association between upstream water availability and the slope of the heat-mortality relationship during the summer. In figure 9, when upstream water availability is above the 25$^{\text{th}}$ percentile, there is no discernible relationship between the mortality rate and the quartile of heat incidence. By contrast, when upstream water availability is critically low, the mean, median, and 75$^{\text{th}}$ percentile of residual mortality rates are both substantially higher during periods of excess heat. The same relationship can be seen in figure 10 with continuous variation in heat incidence rather than quartiles. While there is no discernible difference between the two curves in the absence of excess heat (i.e.,

[7]The coefficient of interest γ_u is still significant and negative when estimating equation 5 on the full sample.

CDD 75°F = 0), the residual mortality rate differentially increases when upstream water availability is low.

Table 12 presents estimates of equations 5 and 6, respectively. As demonstrated in table 12, there is a significant and negative interaction between upstream dam levels and the heat-mortality relationship, including when within-district and downstream dam levels are included as controls.[8] Employing the test of minimum relative selection on unobservables necessary to nullify the effect estimated in column 3 of Table 12 developed by **?**, I obtain a bound estimate of approximately 10.69, substantially exceeding the recommended minimum of 1 (proportional selection). Since dam levels are standardized, the point estimates in Table 12 suggest the slope of the heat-mortality relationship at the mean of upstream water availability (β_{CDD}) is 0.91 deaths per million per CDD 90°F, and a one-standard-deviation increase (decrease) in upstream water availability reduces (increases) this slope by 0.90. In other words, when upstream water availability is one standard deviation above its mean, the slope of the heat-mortality relationship is statistically indistinguishable from zero, with a point estimate very close to zero. In the Online Appendix, Tables 28 through 31 confirm that this estimate is robust to a number of alternative specifications, including using CDD 75°F instead of CDD 90°F to reduce censoring of heat below the extreme threshold of 90°F, using standardized temperature values to account for the possibility of acclimatization to local average temperatures, excluding within-district levels to more closely resemble the analyses in prior literature (e.g. Garg et al. (2018)), and the inclusion of fixed effects for each month in the sample.

Column 4 of Table 12 confirms that the effect of residual upstream water availability on the mortality rate is stronger in the summer. Across all districts, the highest incidence of temperatures exceeding 90°F occurs between October and March, usually peaking in December or January. If water availability moderates the heat-mortality relationship specifically, the effect on mortality should be stronger in months when excess heat occurs more frequently.

[8] In every table reporting regression results, I include specifications that omit downstream dam levels alongside specifications that include it because 3 of the 52 districts are at too low an elevation to have any dams classified as downstream.

This is evident in column 4 of table 12, in which the coefficient on the interaction between upstream dam levels and an indicator for the summer is negative and significant ($p < 0.05$).

2.5 Conclusion

In this paper, I show that investment in water infrastructure can be an effective community-level adaptation to heat. I estimate that a one-standard-deviation increase in residual up-stream water availability makes the slope of the heat-mortality relationship indistinguishable from zero, even in the absence of widespread residential air conditioning. In Online Appendix Section A, I find strongest effects for women, who are most likely to be responsible for re-trieving water for households without a private connection, and infant mortality, which is a natural corollary of the results for women. I also find suggestive evidence of heterogeneity by the location of households' primary water sources, showing that maintaining the historical average level of water availability is sufficient to eliminate the heat-mortality relationship for population groups more likely to have a water source on-premises (e.g. White South Africans), while further increases above the historical average are necessary for those less likely (e.g. Black African and Coloured South Africans). Finally, in Online Appendix Section B, I confirm that upstream water availability is only significant following periods of scarce local precipitation when the local water supply is most likely to be insufficient.

I corroborate these findings in Online Appendix Section C using the Lesotho Highlands Water Project (LHWP) transfer as a natural experiment increasing potable water availability in receiving districts. After the inauguration of the transfer in 2004, minimum dam levels increased and summer mortality rates differentially declined in districts receiving water. This both lends additional credence to the causal interpretation of the panel-fixed-effects estimates in the prior section and demonstrates that the mitigating effect of water availability on the heat-mortality relationship is not conditional on that water being delivered through naturally-occurring rivers and streams. Thus manually increasing water supply and access through technology could significantly reduce the long-run mortality consequences of climate change, as heat waves become more frequent and more intense. I argue this is especially relevant in South Africa because of the projected effects of climate change on rainfall patterns (Allen et al. 2014, Nkhonjera 2017), and as demonstrated in Section B.2, the significance of availability of water from other sources emerges specifically when precipitation is sparse.

As a policy, investment in water access has many potential benefits beyond adaptation. Improved potable water infrastructure has been shown to improve mental health (Devoto et al. 2012), human capital accumulation (Beach et al. 2016), cognitive ability (Troesken et al. 2011), and educational attainment (Ao 2016), as well as agricultural productivity and resilience to rainfall variation (Duflo and Pande 2007) for its beneficiaries. While other adverse effects of construction are possible, especially for households and local ecosystems displaced by infrastructure (Duflo and Pande 2007, Hitchcock 2012, Keketso 2003), investment in water is less likely to significantly increase long-run greenhouse gas (GHG) emissions than encouraging air conditioning (A/C) adoption (Davis and Gertler 2015, Kahn 2016, Wolfram et al. 2012), which has also been shown to weaken the heat-mortality relationship (Barreca et al. 2015, 2016).

This paper adds mitigation of heat-related mortality to the already long list of potential benefits of water, sanitation, and hygiene (WaSH) investment in developing countries. The status quo of WaSH in South Africa leaves much room for improvement, with only 73% of households having access to a safely managed source of water versus about 91% in the OECD (WHO and UNICEF JMP 2015). Investments similar to LHWP to increase water supply can improve access to safe, potable water, while also reducing heat-related mortality risk as global surface temperatures rise. In brief, good WaSH policy is good climate policy in the developing world, and vice-versa.

Future research is needed to identify the mechanisms by which potable water availability mitigates heat-related mortality and their relative importance. More detailed microdata than the panel used in this paper is needed to achieve this. There are several potential channels by which water could interact with heat-related morbidities: drinking water prevents dehydration and lowers the risk of heat stroke; making higher-quality water available reduces the spread of waterborne diseases more likely to be fatal in heat; sanitary water for cleaning can prevent the spread of infectious diseases that spread faster in heat. Determining which of these mechanisms drives the mitigating effect of water availability on heat-related mortality will elucidate which policies and infrastructure projects are most effective to protect vulnerable populations from climate change.

2.6 Figures

Figure 5: Temperature-Mortality Relationship for South Africa, 1997-2016

Figure 6: Map of GHCN-Daily Weather Stations with Non-missing Data, 1997-2016

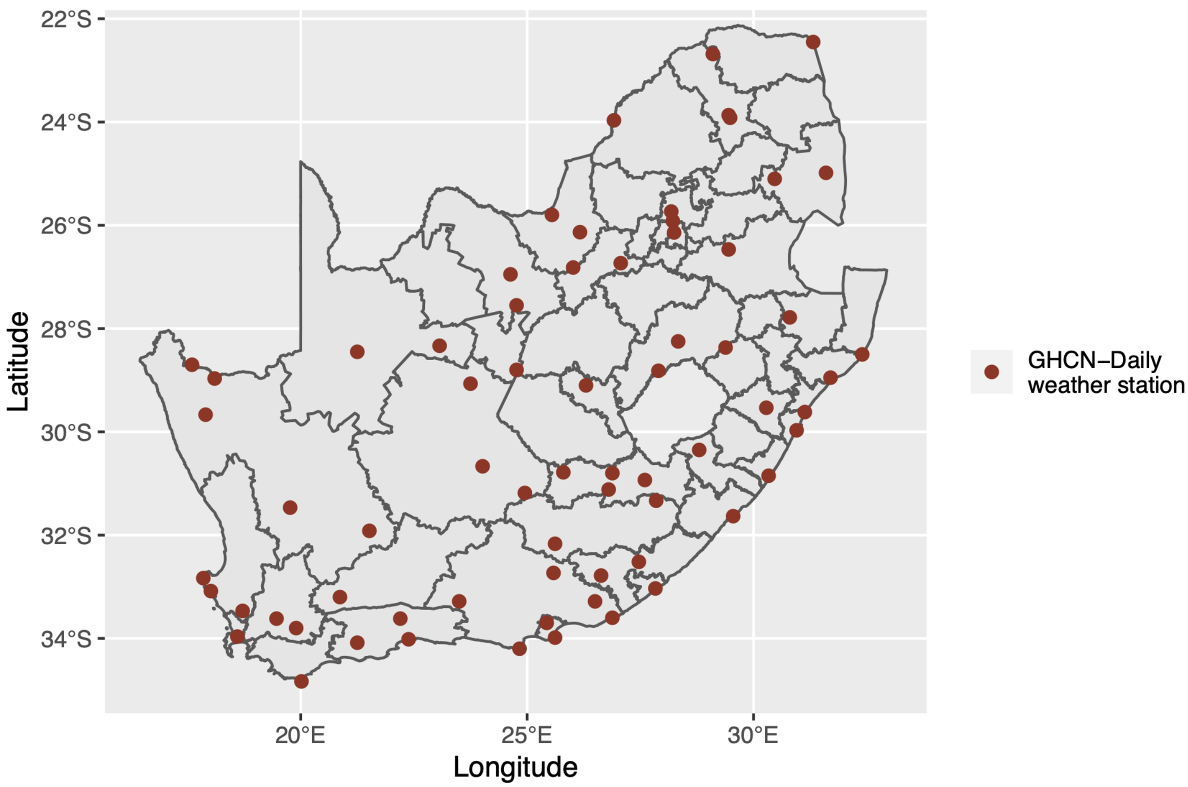


Figure 7: Map of All Monitored Dams, Elevation, and Catchment Areas

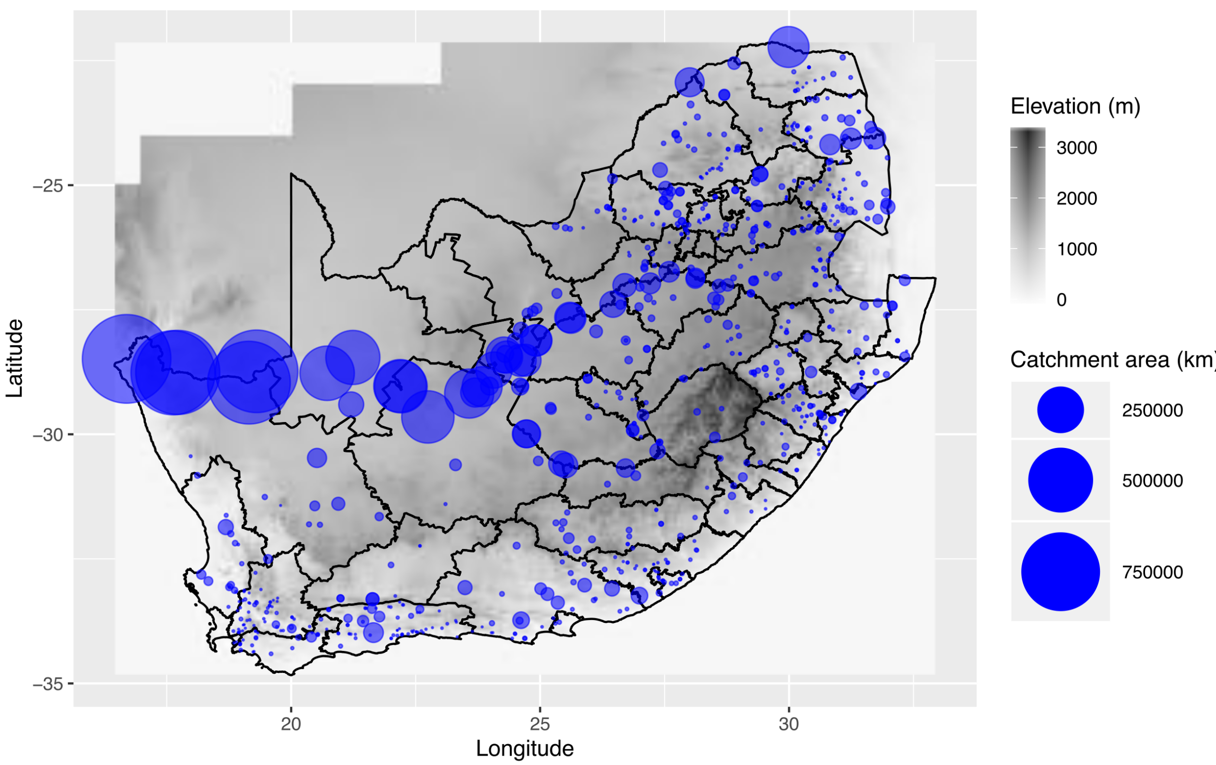


Figure 8: Map of Dams Selected for the City of Johannesburg by Weight

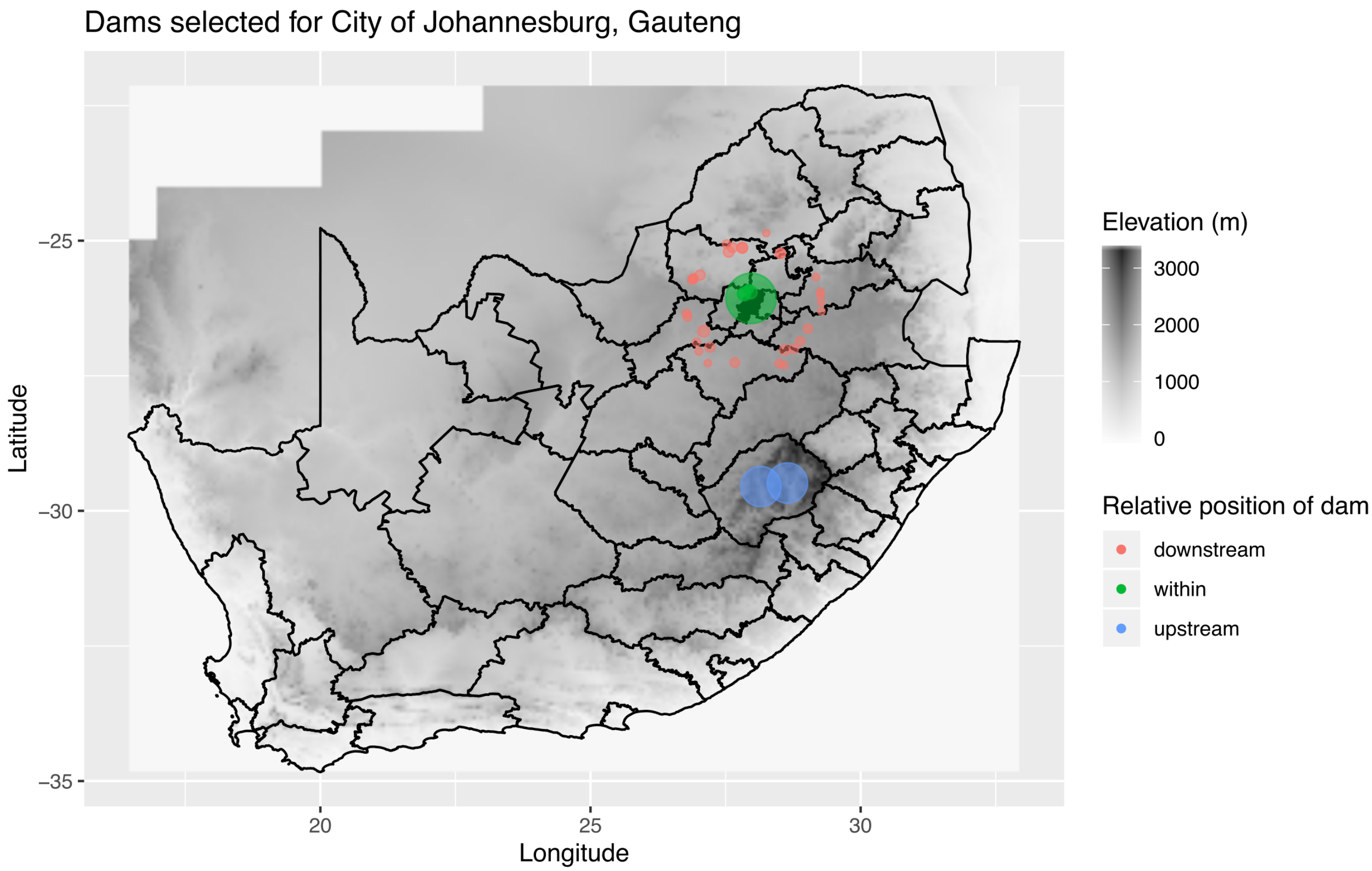
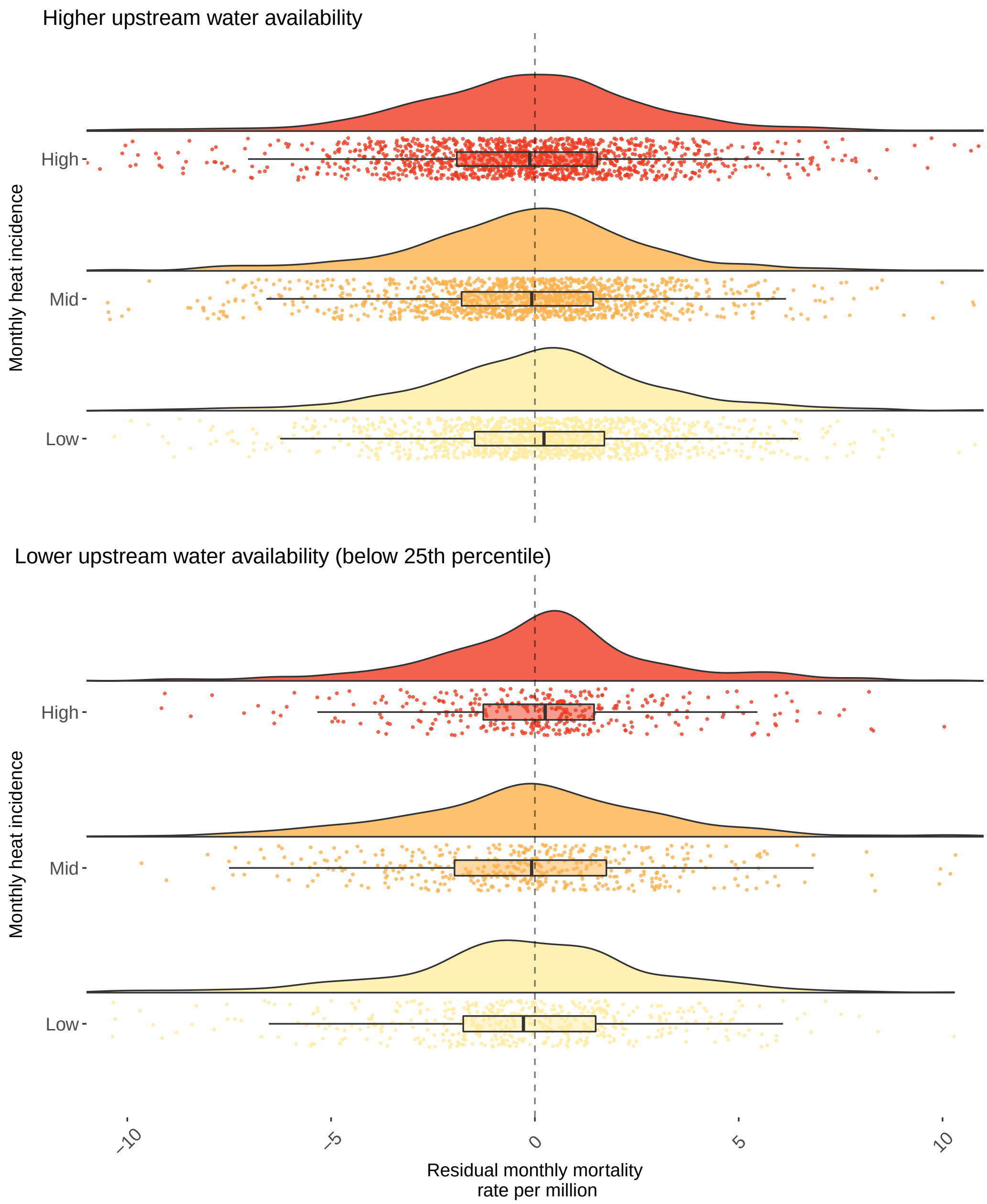
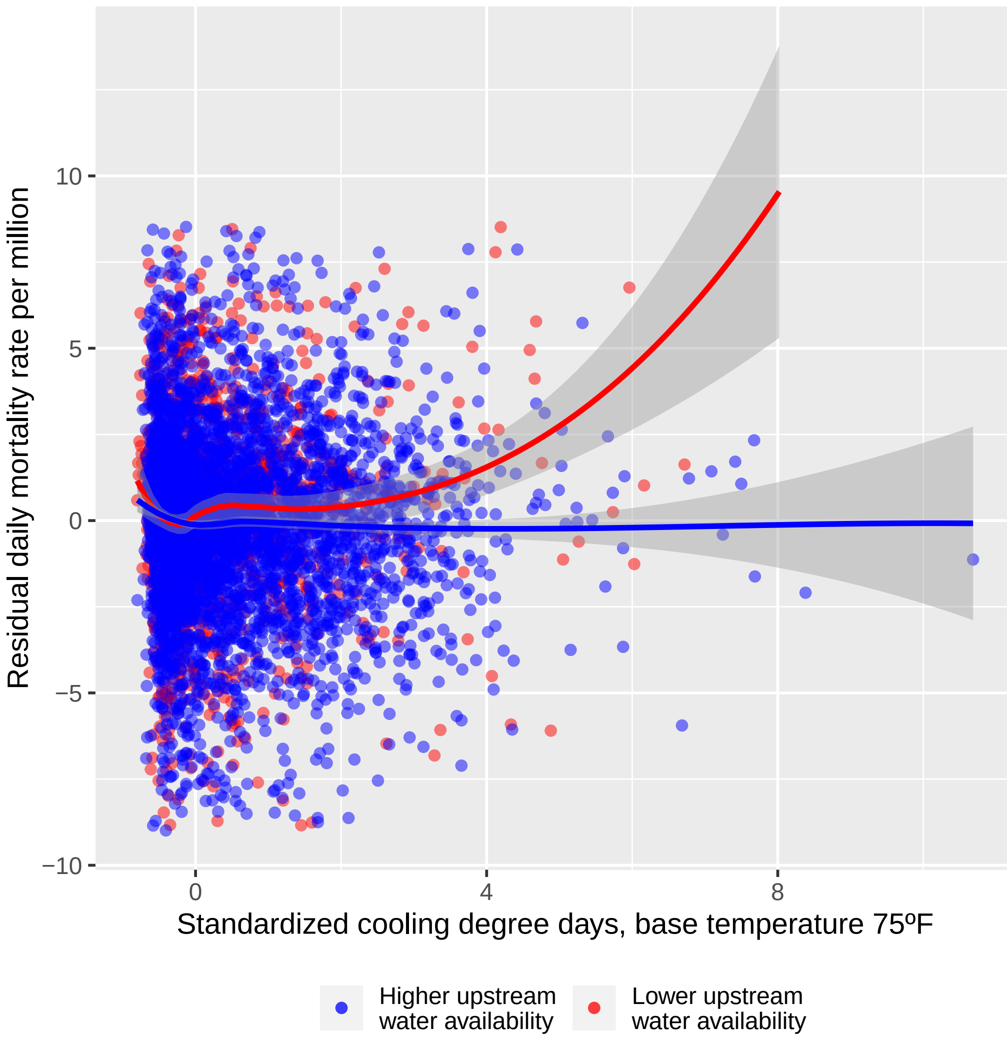


Figure 9: The Distribution of Mortality Rates by Heat Incidence and Water Availability Levels



Note: "High," "Mid," and "Low" heat incidence refer to cooling degree days (CDD) in the top third, middle third, and bottom third of the October to March sample, respectively. "Higher upstream water availability" refers to average upstream dam levels above the 25th percentile.

Figure 10: Local Regression Smoothing (LOESS) of Mortality Rates on Heat Incidence by Water Availability Level



"Cooling degree days, base temperature 75°F" is a measure of duration and intensity of heat exposure above 75°F. For reference, 1 CDD 75°F in a particular month is equivalent to the average outdoor temperature exceeding 75°F by 1 degree for 1 day during that month. In the above figure, this measure is standardized by district to remove baseline geographic differences in climate. "Residual daily mortality rate per million" is the average daily mortality rate per million with district-year fixed effects swept out.

2.7 Tables

Table 11: Summary Statistics by Province

Province	Eastern Cape	Free State	Gauteng	KwaZulu-Natal	Limpopo	Mpumalanga	Northern Cape	North West	Western Cape
Deaths (daily, per million)	35.41 (10.26)	55.15 (51.71)	34.96 (22.01)	32.56 (13.65)	31.23 (18.10)	28.11 (9.30)	33.65 (9.28)	39.72 (14.26)	24.23 (8.22)
Deaths, Oct-Mar (daily, per million)	34.69 (10.21)	52.26 (48.90)	32.45 (20.27)	31.71 (13.08)	29.93 (16.83)	26.96 (8.93)	32.08 (8.76)	37.81 (13.39)	22.83 (7.43)
Male [female] life exp., 2000−2006	51.7 [54.8]	46.5 [49.2]	55.8 [58.6]	48.8 [54.0]	52.0 [55.4]	52.0 [55.6]	52.2 [57.4]	49.9 [54.0]	59.2 [64.1]
Male [female] life exp., 2011−2016	56.1 [62.9]	53.1 [58.8]	62.0 [67.2]	55.3 [61.4]	56.4 [62.8]	57.6 [63.2]	57.2 [63.5]	55.3 [62.8]	63.9 [70.3]
CDD base 75°F (Oct-Mar, monthly)	3.81 (4.62)	6.74 (9.06)	4.02 (5.53)	12.20 (13.04)	12.84 (12.40)	9.53 (13.91)	13.32 (15.27)	34.25 (32.35)	9.43 (10.11)
Mean elevation of dams (m)	584.0 (392.3)	1411.6 (135.6)	1327.5 (122.8)	740.3 (485.5)	838.4 (338.8)	1074.6 (438.4)	900.8 (315.4)	1191.2 (148.5)	305.1 (241.8)
Gastroenteritis deaths	0.04 (0.04)	0.06 (0.04)	0.04 (0.03)	0.06 (0.04)	0.08 (0.05)	0.07 (0.05)	0.04 (0.05)	0.05 (0.04)	0.02 (0.03)
Black African pop, 2015	0.87 (0.34)	0.87 (0.34)	0.80 (0.39)	0.87 (0.34)	0.98 (0.15)	0.93 (0.25)	0.55 (0.50)	0.91 (0.29)	0.35 (0.48)
Head of household age, 2015	48.2 (16.8)	46.7 (15.5)	44.8 (14.1)	46.9 (16.2)	46.5 (16.9)	45.3 (15.6)	48.1 (15.9)	45.7 (15.4)	46.7 (14.5)
Educ $\geq$ HS diploma	0.15 (0.12)	0.18 (0.15)	0.27 (0.18)	0.19 (0.16)	0.14 (0.14)	0.16 (0.16)	0.16 (0.13)	0.15 (0.14)	0.28 (0.17)
Access to piped water	0.42 (0.04)	0.89 (0.03)	0.89 (0.02)	0.54 (0.03)	0.43 (0.03)	0.68 (0.02)	0.79 (0.04)	0.62 (0.02)	0.9 (0.02)

Means unless otherwise specified; standard deviations in parentheses.

Table 12: Heat-Mortality Relationship Above 90°F Interacted with Dam Levels

	Dependent variable: average daily deaths per million			
	(1)	(2)	(3)	(4)
	Oct-Mar	Oct-Mar	Oct-Mar	All months
Upstream dam level × CDD 90°F	-0.87^{***}	-0.94^{***}	-0.90^{***}	
	(0.17)	(0.27)	(0.33)	
Upstream dam level × Summer indicator				-0.78^{**}
				(0.30)
CDD base temp 90°F	0.80^{***}	0.80^{***}	0.91^{***}	
	(0.25)	(0.27)	(0.17)	
Upstream dam level	-1.02^{***}	-0.33	-0.23	0.84
	(0.20)	(0.27)	(0.31)	(0.62)
Within-district dam level × CDD 90°F		0.10	0.13	
		(0.47)	(0.54)	
Within-district dam level × Summer indicator				0.11
				(0.23)
Within-district dam level		-0.95^{***}	-0.82^{***}	-0.80^{**}
		(0.26)	(0.29)	(0.37)
Downstream dam level × CDD 90°F			-0.19	
			(0.42)	
Downstream dam level × Summer indicator				0.88^{***}
				(0.27)
Downstream dam level			-0.33	-1.36^{**}
			(0.38)	(0.54)
Mean of dep. var.	34.12	34.12	34.12	35.37
District-month-of-year FE	No	No	No	Yes
Monthly precipitation control	No	No	Yes	Yes
N	5474	5474	5474	11649
R^2	0.977	0.977	0.977	0.984

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. District-clustered standard errors in parentheses. All columns include district-year fixed effects. "Average daily deaths per million" is the monthly death rate per million divided by the number of days in the month to address the fact that, all else equal, more deaths occur during months with more days. Dam levels are distance-weighted 12-month lagged averages calculated with equations 2 and 3. A dam is classified as upstream (downstream) of a district if it is at least 100 kilometers away from and 100 meters higher (lower) in elevation than the geographic center of the district. A dam is classified as within-district if it is within 100 kilometers and ± 100 meters elevation of the geographic center of the district. For graphical examples, see figures 7 and 8. Wherever within-district or downstream dam level controls are included, their interaction with CDD is also included. Cooling degree days (CDD) are a monthly incidence measure of heat above the base temperature. For reference, 1 CDD at a base temperature of 90°F is equivalent to 1 day in a month during which the average outside temperature exceeded 90°F by 1°F.

3.0 The Regressive Costs of Drinking Water Contaminant Avoidance

Up to 45 million Americans in a given year are potentially exposed to contaminated drinking water, increasing their risk of adverse health outcomes. Existing literature has demonstrated that individuals respond to drinking water quality violations by increasing their purchases of bottled water and filtration avoidance, thereby avoiding exposure to contaminants. This paper demonstrates that poorer households, for whom the costs of avoidance comprise a greater share of disposable income, bear disproportionate costs of water quality violations in the United States. During an active health-based water quality violation in their county of residence, the nutritional content of poor households' purchases from grocery retailers differentially declines by about 22 calories (about 1.8% of the mean) per person per day on average. Event study estimates indicate the effect size increases with the duration of the water quality violation. This finding suggests that the indirect costs of drinking water contamination through economic channels exacerbate health disparities associated with poverty.[1]

3.1 Introduction

According to the Environmental Protection Agency's budget proposal for fiscal year 2020, over 7% of the United States' population served by community water systems receives water that does not meet at least one health-based drinking water standard established by the Safe Drinking Water Act of 1974—at least the ones we know about. Monitoring and enforcing water quality standards is a massive undertaking, requiring periodic sample

[1]Researcher(s) own analyses calculated (or derived) based in part on data from The Nielsen Company (US), LLC and marketing databases provided through the Nielsen Datasets at the Kilts Center for Marketing Data Center at The University of Chicago Booth School of Business. The conclusions drawn from the Nielsen data are those of the researcher(s) and do not reflect the views of Nielsen. Nielsen is not responsible for, had no role in, and was not involved in analyzing and preparing the results reported herein.

collection at the almost 400,000 public water systems registered in the EPA's Safe Drinking Water Information System (SDWIS), testing by certified labs for a broad panel of regulated contaminants, and legal proceedings in the event of an unresolved violation. Given the many layers of coordination required on a national scale, it is unsurprising that an estimated 26-38% of violations are either unreported or erroneously recorded in SDWIS (Allaire et al. 2018, United States Environmental Protection Agency 2000), and many of those which are accurately recorded take several months to return to compliance. According to the EPA, serious health-based violations are "expected to be resolved within six months" (United States Environmental Protection Agency 2020); however, many are not. For example, during the high-profile violation of water quality standards in Flint, Michigan, when the maximum contaminant level for trihalomethanes was exceeded in December 2014, return to compliance was not achieved until 9 months later.[2]

As customers wait for the supplier to fix the problem, either voluntarily or following legal action by the state or federal government, many purchase bottled water in order to avoid contaminants (Allaire et al. 2019, Marcus 2020, Zivin et al. 2011). Expenditure on avoidance methods has been used in the environmental economics literature to estimate the willingness to pay (WTP) for water quality improvements (Brouwer et al. 2015, Brox et al. 2003, Johnston and Thomassin 2010, Rodriguez-Tapia et al. 2017) and similarly for air quality improvements (Freeman et al. 2019). Prior literature using survey-based elicitation methods has estimated average WTP for water quality improvements of between \$5 and \$15 per month in 2020 U.S. Dollars (Brox et al. 2003).

While enforcement of water quality standards is publicly funded, the costs of avoidance are privately borne, and do not scale with income in the absence of targeted subsidies. Thus the need to avoid contaminated water is a regressive income shock. The amount of water required for survival is not a function of income or preferences, so a wealthy household and a poor household with the same number of occupants would need to obtain similar amounts

[2]Perhaps surprisingly, a water quality violation associated with lead levels never appeared in SDWIS, even though the Flint water crisis is most commonly associated with lead in the water. This is likely in part because lead typically enters the household water supply through the corrosion of pipes near the household, rather than contamination occurring earlier in the distribution network.

of bottled water to replace their tap water consumption. However, the cost of this bottled water has a different impact on each household's budget—the wealthy household, unlikely to be budget-constrained, can add bottled water purchases onto their existing consumption patterns, while the poor household is much more likely to need to forego other purchases to fit the cost of avoidance into their budget. Thus poorer households face a difficult trade-off: either maintain current consumption and risk exposure to contaminated water, or purchase avoidance and make sacrifices elsewhere. Even among households who do not purchase bottled water, knowledge of contamination in their tap water may induce other types of substitution with implications for nutritional quality and food security, such as purchasing ready-to-eat foods (including food away from home) to avoid cooking with the water. Calorie-for-calorie, these "convenience foods" are more expensive and of lesser nutritional quality on average (McDermott and Stephens 2010, Rahkovsky et al. 2018, Saksena et al. 2018).

In this paper, I show that during an active water quality violation in SDWIS, the nutritional content of poor households' purchases from grocery retailers differentially declines by about 22 calories per household member per month on average. This effect coincides with an increase in bottled water purchases and a decrease in calories-purchased-per-dollar among the same households. Event study estimates indicate the effect is driven by violations which remain active for longer than 6 months, with significantly larger estimated effect sizes (about 47 calories per person per day, or 3.7% of the mean) beyond month 6 of a long-term violation. Since treatment is assigned at the county level because the water supplier of each individual household is unobserved, these results should be interpreted as intent-to-treat estimates and likely represent a lower bound of the true effect of water quality violations.

This finding contributes to multiple literatures on the economics of natural resources, poverty, and nutrition. Improvements in potable water supply have been linked to increases in happiness (Devoto et al. 2012), decreases in obesity risk (Ritter 2019), decreases in overall mortality risk (Clay et al. 2014, Ferrie and Troesken 2008, Troesken 2004), and increases in human capital accumulation (Ao 2016, Beach et al. 2016, Troesken et al. 2011). These outcomes have also been linked to improvements in nutrition, especially among infants and children (Adhvaryu et al. 2019, Anderson et al. 2016, Frisvold 2015, Hoynes et al. 2011, 2015, Hoynes and Schanzenbach 2009, Kohler et al. 2017). The findings in this paper suggest a potential link between these parallel literatures: in addition to the direct benefits of water supply improvements, poor households may additionally benefit from a positive income effect as improvements to water supply allow them to reallocate their spending on avoidance devices to other nutritious products. Taking this income effect into account increases the long-run value of residential water infrastructure improvements, for which the EPA requested over $2 billion in fiscal year 2020 citing broad prevalence of outdated water systems.

This paper also contributes to the literature on the spatial and intergenerational nature of poverty in the United States. The notion that "zipcode is destiny" (Chetty et al. 2018) may be partially explained by spatial heterogeneity and persistence in the quality of the residential water supply. For several reasons, including the construction of lead pipes in the early 20th century (Clay et al. 2014), the location of cities either upstream or downstream of major sources of pollution (Jerch 2018), and historical to present-day environmental injustice (Schaider et al. 2019, Switzer and Teodoro 2018), water quality violations are not evenly distributed across space. Particularly for children, since many of the known health and developmental conditions caused by contaminated water supply are associated with early-life exposure (Beach et al. 2016, Clay et al. 2014, Valent et al. 2004), living in an area with persistent water quality violations increases the risk of conditions that in turn reduce the likelihood of getting out of poverty. If poor water quality requires families to choose between avoidance and adequate nutrition, children are placed in a lose-lose situation, as either choice is likely to have negative consequences for their development. While this paper focuses on the short-term impacts of water quality violations to limit threats to causal identification, the results are consistent with correlational disparities in food security based on an area's history of water quality violations (see Figure 1).

The paper proceeds as follows. Section 2 describes the data used to construct the water quality violation and household consumption measures. Section 3 describes the results. Section 4 discusses the limitations and areas for further research. Section 5 concludes.

3.2 Data and Empirical Strategy

3.2.1 Data

To assess the differential effect of water quality violations on poor households' budget constraints and resulting nutrition, this paper combines household-level panel data on grocery and department store purchases, nutritional information on grocery products, and administrative records of water quality violations and enforcement activities. Each data source is described below.

Household consumption data. Measures of monthly household consumption were constructed from the Nielsen HomeScan Consumer Panel, a nationally representative panel of households' retail purchases, from 2004 to 2016. The panel contains 168,772 unique households spanning 2,967 counties in the United States, and includes purchase dates, quantities, and prices paid for about 2.2 million unique Universal Product Codes (UPCs) for grocery products. Among these UPCs, Nielsen provides "extra attributes" such as flavor, organic labeling, and container type for about 1.3 million. The demographic information provided about each household includes number of residents, annual income, and the gender, race, age, and occupation of each resident. In all analyses reported in this paper, households were classified as "poor" if their reported income in a given year was less than 200% of the established federal poverty line in that year based on the household's number of residents, including children.

UPC-level nutrition facts data. The UPC-level consumption data in the Nielsen HomeScan Consumer Panel was merged with the Nutritionix Consumer Packaged Goods database purchased from Syndigo Inc. This data includes all information from a product's legally mandated nutrition facts label, including calories, fat, carbohydrates, sugars, and protein per serving, and the number of servings per container. Among the UPCs represented in the Nielsen data, 305,455 products were directly matched with a product in the Nutritionix data. To assign nutrition facts to unmatched UPCs, I conduct a two-stage imputation process. In the first stage, I use a semantic matching technique based on the "extra attributes" Nielsen provides for about 1.3 million products to associate each unmatched product with the most similar matched product.[3] Products are semantically matched based on the longest common substring (LCS) distance between attribute strings, which were constructed by concatenating all available attribute description fields.[4] In the second stage, I compute the median value by category of each nutrition facts variable among all (directly or semantically) matched products using 679 "product modules" (e.g. "Ice Cream - Bulk," "Vegetables - Greens - Canned", etc.) pre-defined by Nielsen, and assign these median values to all remaining unmatched products. To ensure comparability across products of different sizes, all imputation is based on standardized per-100-gram measures of nutritional content; after imputation, the total nutritional value for each product is computed by scaling up these per-100-gram values using the provided metric weight of that product.

Water quality violations and enforcement data. The incidence, timing, and type of water quality violations were determined at the county level from the EPA Safe Drinking Water Information System (SDWIS) database, an administrative dataset that records public water system facility locations and populations served, site visit logs, and various types of water quality standard violations from 2009 onward. Health-based violations, which are failures to adhere to established maximum contaminant levels (MCLs) for regulated contaminants such as lead, arsenic, and nitrates or related treatment protocols, are the focus of this paper. SDWIS also separately categorizes some health-based violations as acute when the violated MCL poses an immediate threat to customers' health.

I classify household i in county j as under an "active violation" in month t if there exists a health-based water quality violation recorded in SDWIS for a water supplier that

[3] A similar approach to match nutrition facts to a broader database of products was used in Carlson et al. (2019).

[4] An example of an attribute string is `PIECE MILK CHM MM SMORES CHOCOLATE CHOCOLATE COVERED PIECE` for a S'mores flavored milk chocolate piece product, which includes its product type, flavor, and common consumer name.

serves at least 500 customers[5] in county j for which public notification has been requested by state or federal authorities prior to month t and compliance has not yet been achieved by month t. Because each individual household's water supplier is not observed, there is uncertainty regarding whether or not household i is actually affected by the violation. The results should thus be interpreted as "intent-to-treat" rather than the "treatment on the treated." Because entirely unaffected households should not respond to these violations, this attenuates estimates, and thus the estimated effects in this paper should be considered a lower bound.[6]

To produce the results reported in Section 3.3, I restrict the sample in two ways. First, I only include households that experience at least one health-based water quality violation during the sample period, which comprise about 39% of the overall consumer panel sample. As is evident in Figure 11, many counties do not have any active violations during the sample period. Those that do are a selected subsample of counties that are more rural and lower-income than average, meaning that comparisons along the extensive margin of water quality violations may be confounded by unrelated differences, especially if those differences are time-varying. Instead, this sample restriction means that the differential timing of water quality violations across counties is used to identify the effect of those violations on household consumption. Secondly, I only count violations as "active" if they have *both* relevant dates recorded in SDWIS (the date public notification was requested and the date of return to compliance). This is necessary because it is not possible to differentiate between presently ongoing violations and missing data.

[5]This eliminates "Very Small" water suppliers, which include restaurants, office buildings, and other public institutions that are required to report to state or federal water quality monitors but do not provide water to households on a large enough scale for households in the panel to plausibly be affected. This cutoff was also used in Allaire et al. (2018), citing different reporting requirements and less reliable data from "Very Small" systems.

[6]It is possible for households which are not directly affected by a water quality violation to respond if they hear about a violation in their area and thus become more skeptical of their own water supply. This would simply be a different channel by which local water quality violations affect household consumption with similar consequences.

3.2.2 Empirical Strategy

3.2.2.1 Panel Fixed-Effects Regression

Using the above data sources, the following panel fixed-effects model was estimated:

$$CaloriesPerPersonPerDay_{it} = \delta_1 ActiveViolation_{it} \times Below200PctFPL_i + \beta_1 ActiveViolation_{it} + \beta_2 Below200PctFPL_i i + t + \epsilon_{it} \tag{7}$$

where $Below200PctFPL_i$ is a dummy that equals 1 when the household's total income does not exceed 200% of the corresponding year's federal poverty line based on the household's number of residents, $ActiveViolation_{it}$ is a dummy variable that equals 1 when household i is under an active water quality violation defined above, i and t are vectors of fixed effects for each household i and time period t respectively, and ϵ_{it} is an error term. Standard errors are clustered at the county level because the treatment varies at the county level. To address growing concerns of bias in two-way fixed effects (TWFE) linear regressions (Roth et al. 2022) in a way feasible to implement with high-dimensional data, I apply the weighting method proposed by Imai and Kim (2019).[7]

To further investigate the potential causes and consequences of differential changes in the nutritional makeup of grocery purchases between wealthier and poorer households, the model in equation 7 is re-estimated with two other dependent variables of interest: reported expenditure on bottled water, and reported overall grocery product expenditure. I deflate both of these expenditure measures to 2004 dollars using the annual food-at-home CPI from USDA (Kuhns et al. 2015). Based on prior literature, I hypothesize that bottled water purchases will increase among affected households following public notification of a water quality violation (Allaire et al. 2019, Marcus 2020, Zivin et al. 2011). For budget-constrained households, this may result in a decrease in calories purchased if bottled water expenditure crowds out expenditure on more nutrition-dense grocery products. In accordance with this, I hypothesize that the effect on total expenditure on grocery products excluding bottled water

[7] This was done with the assistance of Laurent Bergé, author of the `R` package `fixest` (Bergé 2018).

will decline among poorer (and thereby more likely budget-constrained) households during an active water quality violation.

Because the treatment indicator $ActiveViolation_{it}$ turns on and off over time within each unit, a key identifying assumption in this specification is that the effect of a water quality violation on household consumption is exclusive to the periods in which that violation is active. In practical terms, this means that households trust that their water is safe to consume immediately after being notified that the violating water supplier has returned to compliance. If this assumption is violated, the estimates generated with this specification are likely to be biased toward zero, since periods in which the treatment indicator is turned off after a violation returns to compliance will still be affected by the violation. Indeed, both anecdotal and empirical evidence from the Flint, Michigan water crisis suggest this assumption may not hold, particularly for severe, high-profile violations; a report by *CNN* in April 2018 highlighted that many residents of Flint, Michigan were still avoiding their tap water despite the government's declaration that it was safe to drink (Chavez 2018), and recent work by Christensen et al. (2019) corroborates this by finding that home prices in Flint remained depressed as a result of the water crisis as late as August 2019.

3.2.2.2 Event Study

To complement the panel fixed-effects regression described in section 3.2.2.1, I conduct an event study which estimates effects of water quality violations on calories per person per day, bottled water expenditure, and total non-bottled-water grocery expenditure for each month relative to the nearest violation's start date. This allows for the visualization of heterogeneous treatment effects over time (e.g. more intense effects for violations that take longer to return to compliance) as well as the assessment of parallel trends pre-violation, which is necessary to confirm that the panel fixed-effects regression results are not driven by the continuation of differential trends pre-treatment that would not be absorbed by unit or time fixed effects. I plot event study coefficients for the 24 months prior and 24 months after each violation, binning all periods more than 25 months before and 25 months after the start date respectively. Like the panel fixed-effects regressions described in the previous subsection, the event study regressions use the weights proposed by Imai and Kim (2019) to mitigate bias in two-way fixed effect linear regressions.

In the event study, period 0 represents a month in which public notification of a violation was requested.[8] All other treatment periods are defined as follows: if there is an active violation in the household's county of residence in a given month, the household's treatment period in that month is the number of months that have elapsed since public notification of that violation was requested; otherwise, the household's treatment period in that month is the (negative) number of months until the *next* violation occurs in that household's county of residence.[9] Much like the panel fixed-effects results reported in Section 3.3, the key identifying assumption of this event study specification is that water quality violations only affect the outcome of interest while they are active.

3.3 Results

Figure 11 maps the number of health-based violations and the food insecurity rate in 2017, as measured in Feeding America's 2019 *Map the Meal Gap* report. The average food insecurity rate is positively correlated with the number of health-based violations since 2010, from an average of 12.7% among counties with no recorded violations to 14.2% for counties with 3 violations or more (the 90th percentile), and this relationship is statistically significant ($p < 0.01$). A similar pattern holds for the child food insecurity rate, for which the average among counties with no violations is 18.6% and the average among counties with 3 violations

[8] Without data limitations, it would be preferable to set period 0 as the month in which public notification was *issued*. However, unfortunately, while some violations do have this date recorded in SDWIS, it is missing for the vast majority of recorded violations.

[9] For example, suppose a household's county of residence is notified of a water quality violation for the first time in January 2015, the associated water supplier returns to compliance in June 2015, and then commits another violation about which the public is notified in September 2015. Then for each month between December 2014 and October 2015, the household is assigned the following treatment periods: -1, 0, 1, 2, 3, 4, 5, -2, -1, 0, 1.

or more is 20.9% ($p < 0.01$).

While this strong positive correlation between food insecurity and water quality violations motivates the research question of this paper, there are many reasons it may not be causal. Both the history of water quality violations and the rate of food insecurity are correlated with income-based poverty rates and themselves could be considered dimensions of a broader definition of poverty. Thus this correlation could be attributable to many other things associated with poverty, including the possibility that wealthier households move away from or avoid moving to an area with a history of water quality violations. To causally identify the effect of water quality violations on household budget constraints and resulting consequences for nutrition, it is necessary to observe consumption within a household before and after a water quality issue emerges.

Table 13 reports the results of the panel fixed-effects regression which estimates the differential effect of an active water quality violation for households below 200% of the federal poverty line relative to their wealthier counterparts. The interaction coefficient in Column 1 of Table 13 is positive and statistically significant, confirming that poor households are responding to county-level water quality violations by increasing their purchases of bottled water.[10] While the point estimate is negative, the interaction coefficient in Column 2 is not statistically significant, suggesting that monthly non-bottled-water expenditure among poor households does not significantly *differentially* decline during a water quality violation. However, the non-interaction term in Column 2 is negative and statistically significant, suggesting that non-bottled-water grocery expenditure declines for all households during a water quality violation. Nonetheless, the nutritional consequences of these consumption differences—the key result of this paper—are substantially larger for poorer households, as demonstrated in Columns 3 through 7. During an active water quality violation, calories per person per day from purchased grocery products differentially declines among poor households by about 22 (1.8% of the mean, $p < 0.05$), with proportionally similar declines across a range of nutrient

[10]The statistically insignificant coefficient on bottled water sales for households above 200% of the federal poverty line is consistent with the finding in Allaire et al. (2019) that the bottled water consumption of higher-income, nonrural counties respond the least to water quality violations.

types. Given the coarse county-level assignment of water quality violations in the data used in this paper, these results should be interpreted as intent-to-treat estimates, and are thus likely significantly biased toward zero.

Figures 12, 13, and 14 present the event study coefficients for calories per household member per day, bottled water expenditure, and total non-bottled-water expenditure, respectively. The event studies for both calories per household member per day and total non-bottled-water expenditure demonstrate no significant pre-trend and effect sizes that increase significantly over the course of a long-lasting violation. While just under half of water quality violations return to compliance within 6 months, with a mode of 1 month (see Figure 15 for a histogram), almost half (47%) are not, and these long-term violations may be especially problematic for poor households. There are multiple potential explanations for this: longer-term violations may receive more coverage, leading to higher awareness of the water quality problems and thus stronger incentives to avoid exposure, especially if legal action is taken against the water system as in Flint, Michigan; longer-lasting violations may mean greater cumulative exposure to contaminants, creating health problems in affected customers that cannot afford avoidance, increasing medical expenses and sick time away from work; or longer-lasting violations may culminate in other forms of economic disruption (e.g. business closures) that effectuate disproportionate negative income shocks on poorer households. More research with more granular data is needed to determine the underlying mechanism of the observed effect heterogeneity with certainty. By contrast, the event study for bottled water expenditure appears flat over the course of a violation and potentially influenced by a positive pre-trend. The implications of this result will be discussed in more detail in the next section.

Finally, Table 14 explicitly tests the time-based heterogeneity suggested by the event studies by splitting the treatment variable into two indicators: one which turns on during months 1 through 6 of a water quality violation, and another which turns on during months 7 and beyond. For this table, I restrict the sample to the households which experience at least one long-term (> 6 month) violation during the study period, which comprise about 54% of households which experience a violation of any duration.[11] This is done to rule out potential confounding of duration-based heterogeneity by unobserved differences between households which experience long-term violations and those which do not. The estimated effect size for months 7 and beyond of an active water quality violation for calories (a decrease of about 47 per person per day, or 3.7% of the mean) is about 2.7 times larger than the estimate for months 1-6, and this difference is statistically significant ($p \approx 0.038$). However, the increase in expenditure on bottled water does not significantly differ between months 1-6 and month 7+ respectively, and while the decrease in total expenditure becomes statistically significant for month 7+, the difference in interaction coefficients is not statistically significant at conventional levels ($p \approx 0.15$).

3.4 Discussion

While the data used in this paper offers many unique advantages, including the ability to analyze within-household effects of water quality violations, broad spatial coverage of the United States, and rich product-level data on household consumption, it also comes with limitations. Several of these limitations have already been discussed, including ambiguity over the exact time in which a water quality violation could be expected to change household behavior, the need to impute nutrition values for the majority of purchased products either through semantic matching or the use of category medians, and the inability to determine for sure whether or not a particular household receives water from a source that committed a quality violation. Further research with more precise and granular data is needed to confirm these findings. This is especially true of the effect magnitudes, since most of the limitations introduce uncertainty that is likely to bias the estimated effects toward zero.

This paper also cannot fully address what households are spending money on instead of nutritive products when a water quality problem emerges. While I do find evidence of in-

[11] A full sample version can be found in Appendix Table 34.

creased bottled water purchases during an active water quality violation, the estimated effect is too small to fully explain the observed concomitant decrease in nutritive purchases. There are several possible reasons for this: contaminated water may cause health problems requiring expensive treatment, especially if the impacted poor households are uninsured; reduced trust in household tap water may discourage preparing foods in the home, thus encouraging households to substitute groceries for fast food and restaurant takeout that would not be represented in the HomeScan Consumer Panel; or households may be purchasing avoidance in ways that are not reported as consumption in the panel, by installing water filters in the home or purchasing water from retailers other than grocery or department stores. The exact long-run health consequences of the effects presented by this paper depend on which explanation holds for the majority of households.

Further research is also needed to determine the effect of water quality violations on *consumed* calories and other macronutrients rather than purchased quantities at grocery stores. The United States Department of Agriculture estimates total food waste in the U.S. between 30 and 40% of the food supply. This suggests that on average, calories consumed will be a proper subset of calories purchased. At the same time, according to USDA ERS, half or more of the food consumed in the United States is prepared outside of the home (e.g. from restaurants) (Saksena et al. 2018) Because the data used in this paper does not include purchases from restaurants, I cannot observe whether households are engaging in substitution of food away from home for groceries (perhaps to avoid cooking with contaminated water) or decreasing their overall food consumption. While food away from home, especially from fast food establishments, tends to be of lower nutritional quality than food prepared at home (Saksena et al. 2018), this means that more complete food consumption data would be required to determine which households or particular household members are facing hunger versus those that are consuming more unhealthy foods.

Finally, the Nielsen HomeScan Consumer Panel data is self-reported by participating households, and thus may feature reporting errors and lapses in participation (Einav et al. 2010). The data used in this paper is restricted to the "purchases static" created by Nielsen, which restricts the sample to households who report their purchases for at least ten months of a given year. Nonetheless, this does not guarantee that households record all of their purchases, and households entering and exiting the panel over time as violations occur may confound the estimated effects. To address this concern, in the appendix, I report results based on the Nielsen Retail Scanner data, which comprises all purchases made at 35,000 participating retailers across the United States. This approach is described in more detail in Appendix Section C.1.

3.5 Conclusion

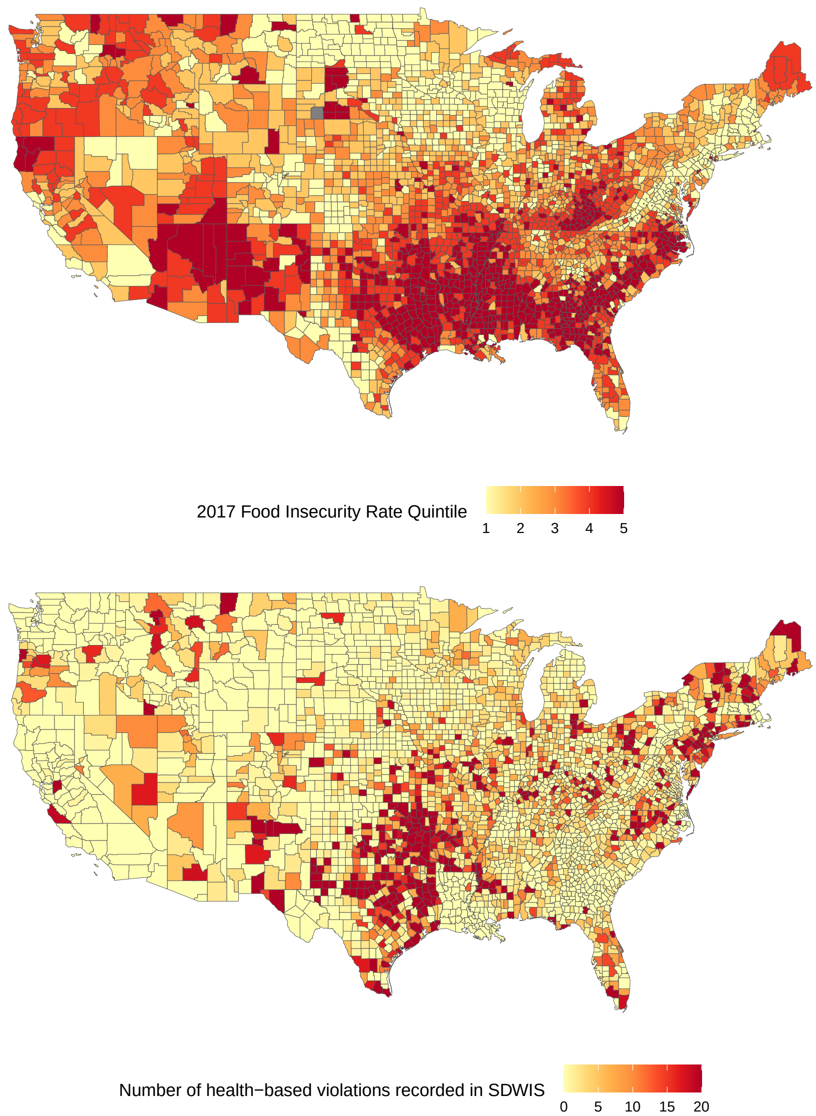
Using a panel of household retail purchases, nutritional quantities based on UPC-level nutrition facts, and county-level water quality violations, this paper demonstrates a negative effect of water supply contamination on household nutritional consumption for households with income below 200% of the federal poverty line. During an active water quality violation, calories purchased per household member per day differentially declines among poorer households by about 22 (1.8% of the mean). Event study coefficients suggest that this effect accumulates over the duration of water quality violations that take several months to return to compliance, with significantly larger estimated effect sizes (about 47 calories per person per day, or 3.7% of the mean) beyond month 6 of a long-term violation. In the same time period, demand for bottled water, which facilitates avoidance of contaminated water, increases among poorer households. By virtue of their lower income, comparable absolute increases in demand for avoidance comprise a larger share of the poorer households' total monthly expenditure. However, observed demand for avoidance cannot account for the entire decrease in nutritive purchases, suggesting a change in preferences (such as buying more food away from home to avoid cooking with the contaminated water) or unobserved avenues of purchasing avoidance.

These findings suggest multiple potential welfare gains from increased investment in water quality monitoring, enforcement, and improvement. An estimated 26-38% of water quality violations are either incorrectly or never reported to the federal Safe Drinking Water Information System (SDWIS) (United States Environmental Protection Agency 2000). Since the findings of this paper focus on the time leading up to the documentation of a water quality violation rather than the actual public notification of the violation, this suggests that households in areas with persistent undetected water quality violations may experience long-term costs that undermine nutrition and other welfare-improving expenditures without recourse. Even among the violations that are detected, significant delays are likely because testing is costly and there are a large number of water suppliers to monitor; the longer these delays are, the more costs accumulate to affected households. Finally, improvements to residential water supply infrastructure would reduce the likelihood of certain types of violations such as dissolved lead regardless of monitoring frequency. In addition to the many known direct benefits of improved water supply (Ao 2016, Clay et al. 2014, Devoto et al. 2012, Ferrie and Troesken 2008, Ritter 2019, Troesken et al. 2011), this paper adds reduced risk of food insecurity to the list of potential welfare improvements resulting from these investments.

This paper also provides additional evidence of spatially-determined disparities in the United States. Much of the residential water infrastructure in centuries-old U.S. cities is outdated, hence the EPA's request of more than $2 billion in fiscal year 2020 to fund reconstruction projects nationwide. The quality of that infrastructure thus to some extent reflects the historical prosperity of each particular neighborhood. If that infrastructure increases the likelihood of repeated or persistent water quality violations, it will inhibit economic development, as both people and businesses are discouraged from moving in to an area with unsafe water; this in turn makes it more difficult to replace that infrastructure without public funding. Meanwhile, this paper suggests those who live in that neighborhood must choose between exposure to harmful water contaminants or foregoing food on the table to purchase avoidance. Both options carry a risk of developing health conditions, especially for children, that increase the difficulty of social mobility.

3.6 Figures

Figure 11: Maps of 2017 Food Insecurity Rate and Health-Based Violations Since 2010



Note: A health-based violation is included in a county's total if it affects at least 500 customers, based on the population served indicated in SDWIS.

Figure 12: Event Study of Effect on Total Calories Purchased Per Household Member Per Day

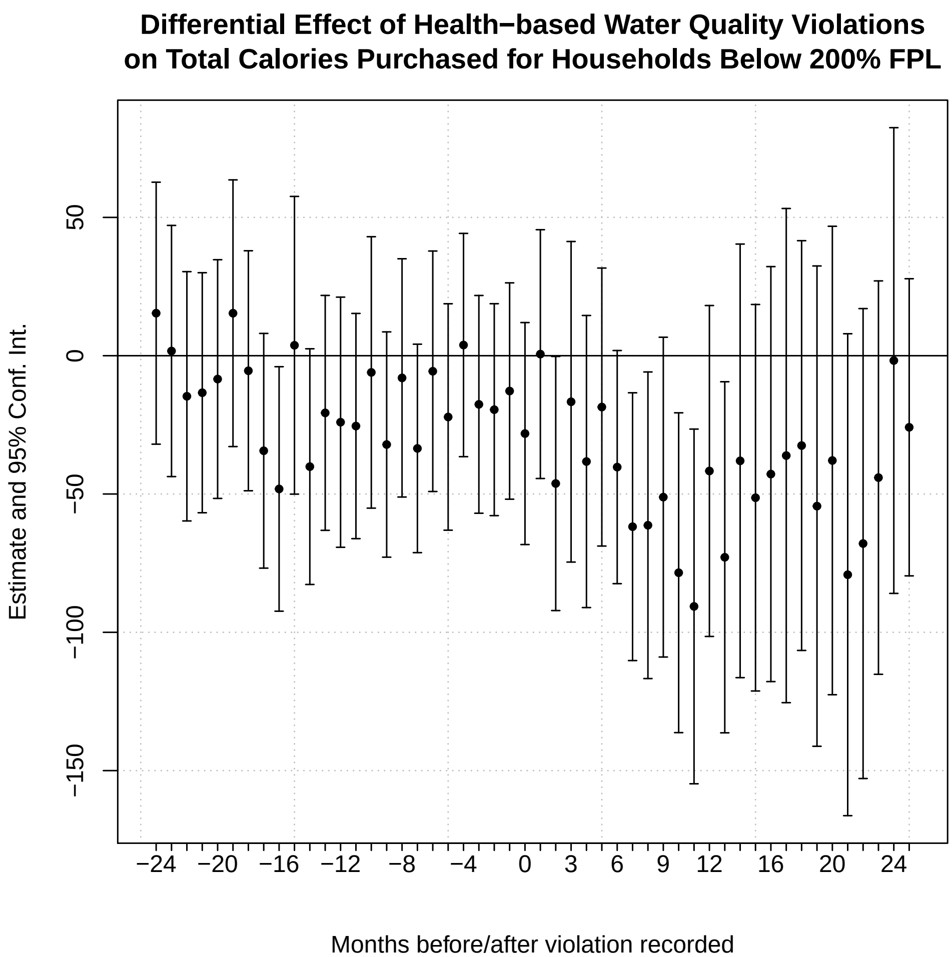


Figure 13: Event Study of Effect on Monthly Household Expenditure on Bottled Water

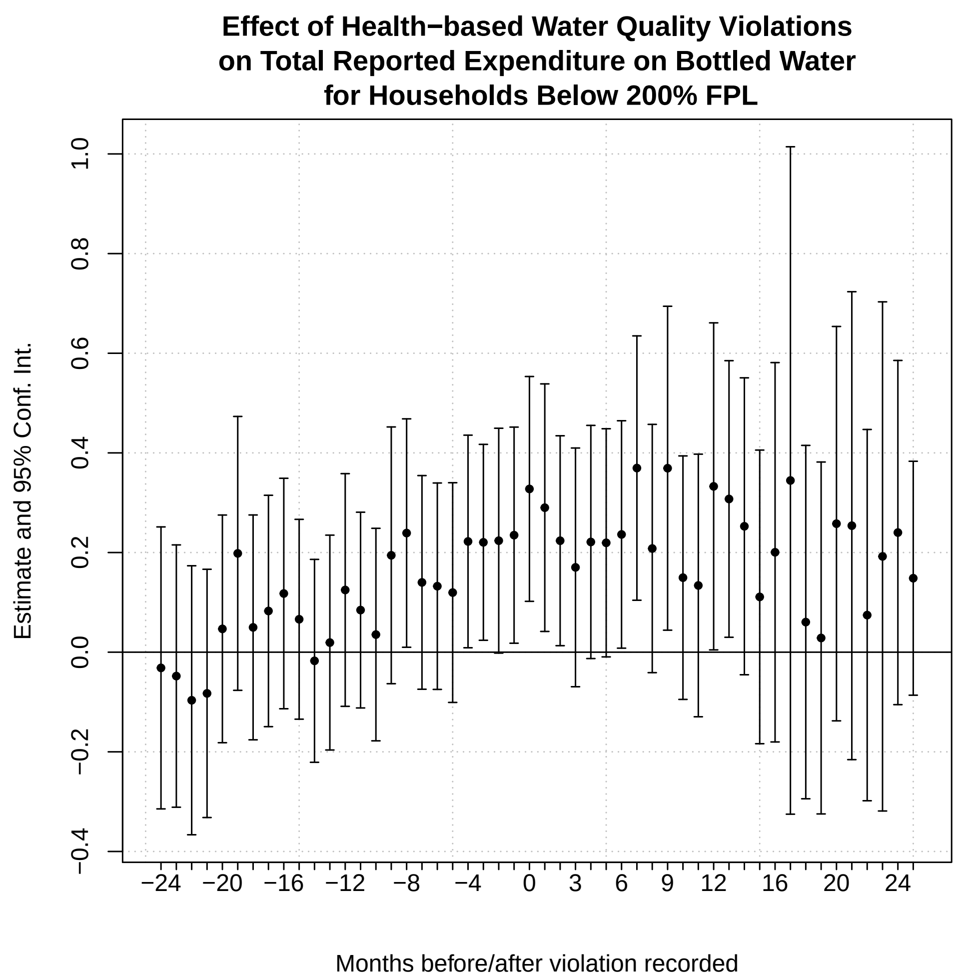


Figure 14: Event Study of Effect on Total Monthly Expenditure

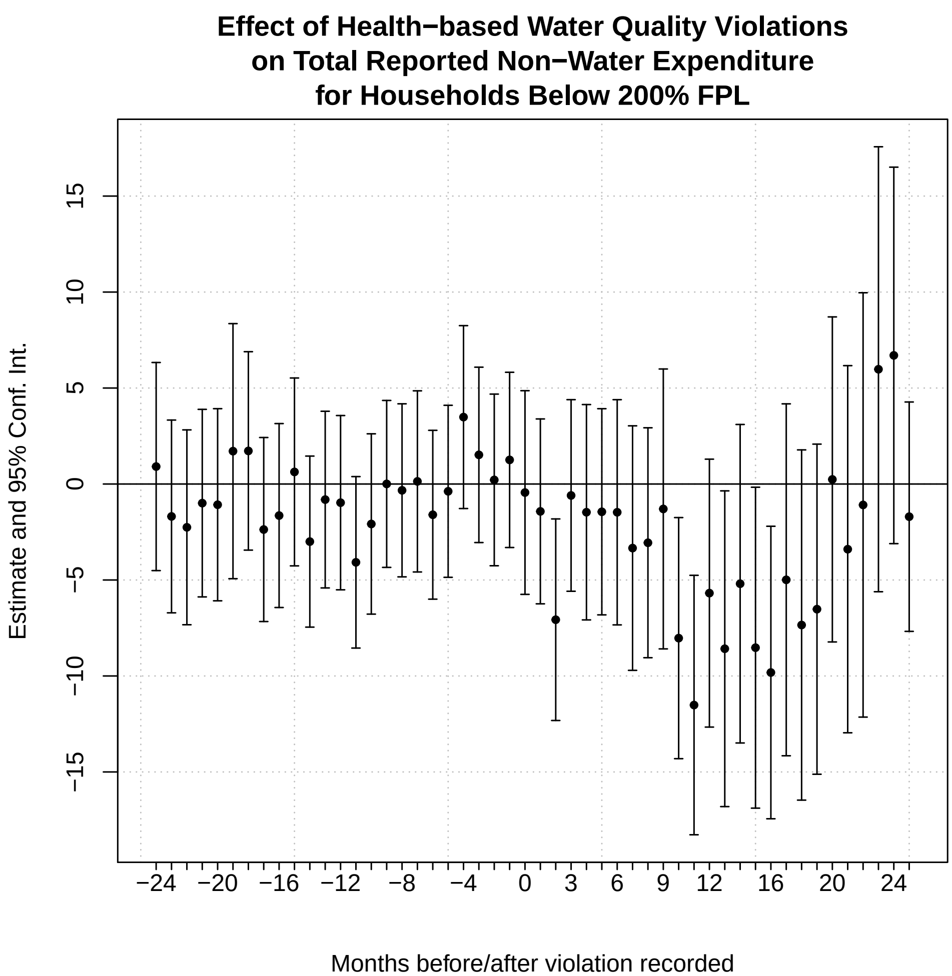
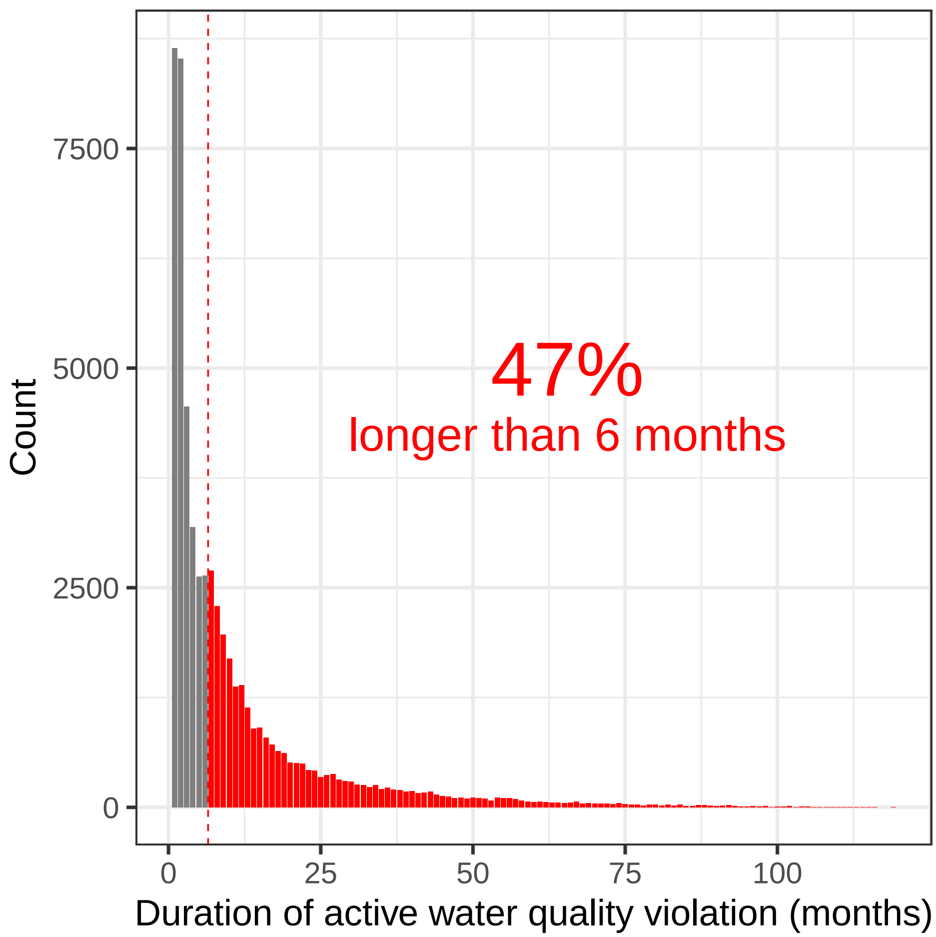


Figure 15: Histogram of Active Water Quality Violation Durations



3.7 Tables

Table 13: Differential Effects of Active Health-Based Water Quality Violations

	Bottled water exp. (2004 USD)	Other exp. (2004 USD)	Calories (Cal)	Fat (Grams)	Carbs (Grams)	Sugars (Grams)	Protein (Grams)
Active Violation ×	0.136***	-0.523	-21.916**	-0.970**	-2.499**	-2.001***	-0.782***
Below 200% FPL	(0.048)	(1.824)	(9.119)	(0.443)	(1.236)	(0.769)	(0.296)
Active Violation	-0.002	-0.860**	-4.913**	-0.167	-0.524**	-0.277*	-0.138**
	(0.012)	(0.428)	(2.159)	(0.104)	(0.265)	(0.155)	(0.064)
Below 200% FPL	-0.055	-3.925***	-89.615***	-3.777***	-11.318***	-5.022***	-1.980***
	(0.048)	(1.392)	(11.297)	(0.498)	(1.382)	(0.676)	(0.329)
Mean of dep. var.	2.03	204.30	1233.78	50.96	152.66	68.19	32.14
Coef. % of mean	6.7%	0.02%	1.8%	1.9%	1.6%	2.9%	2.4%
N	4328963	4328963	4328963	4328963	4328963	4328963	4328963
Adjusted R^2	0.374	0.564	0.430	0.417	0.390	0.313	0.438

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. County-clustered standard errors are in parentheses. Expenditure measures were deflated to 2004 USD using the annual food-at-home CPI measures from USDA (Kuhns et al. 2015). Household and panel month fixed effects are included in all columns. Regressions were weighted using the method proposed by Imai and Kim (2019) to mitigate bias in two-way fixed effect (TWFE) linear regressions.

Table 14: Differential Effects of Active Health-Based Water Quality Violations Based on Duration (Restricted Sample)

	Bottled water exp. (2004 USD)	Other exp. (2004 USD)	Calories (Cal)	Fat (Grams)	Carbs (Grams)	Sugars (Grams)	Protein (Grams)
Active Violation (Month 7+) × Below 200% FPL	0.169***	-3.839**	-46.569***	-1.657**	-5.627***	-3.218***	-1.172**
	(0.064)	(1.773)	(14.309)	(0.682)	(1.818)	(1.008)	(0.471)
Active Violation (Month 7+)	0.006	0.170	-4.237	-0.209	-0.341	0.084	-0.243**
	(0.018)	(0.648)	(3.430)	(0.154)	(0.435)	(0.242)	(0.106)
Active Violation (Month 1-6) × Below 200% FPL	0.201***	-1.469	-17.175	-0.355	-1.781	-1.285	-0.695*
	(0.064)	(1.575)	(12.324)	(0.582)	(1.725)	(1.048)	(0.375)
Active Violation (Month 1-6)	0.002	-0.464	0.821	0.062	0.174	-0.009	-0.033
	(0.015)	(0.760)	(2.685)	(0.125)	(0.323)	(0.186)	(0.078)
Below 200% FPL	-0.115**	-1.915	-80.755***	-3.674***	-10.435***	-4.545***	-1.838***
	(0.053)	(1.470)	(14.621)	(0.651)	(1.784)	(0.902)	(0.422)
Mean of dep. var.	2.05	204.30	1258.37	52.22	155.15	76.66	31.74
P-value: Short-term vs. long-term	0.594	0.161	0.038	0.046	0.042	0.079	0.275
N	2469983	2469983	2469983	2469983	2469983	2469983	2469983
Adjusted R^2	0.356	0.549	0.412	0.399	0.372	0.294	0.422

Notes: * p < 0.1, ** p < 0.05, *** p < 0.01. County-clustered standard errors are in parentheses. “Restricted Sample” refers to the subset of households who experience at least one long-term (> 6 month) water quality violation during the sample period. Expenditure measures were deflated to 2004 USD using the annual food-at-home CPI measures from USDA (Kuhns et al. 2015). Household and panel month fixed effects are included in all columns. Regressions were weighted using the method proposed by Imai and Kim (2019) to mitigate bias in two-way fixed effect (TWFE) linear regressions. The row “P-value: short-term vs. long-term” reports statistical comparison tests between the coefficient for “Active Violation (Month 7+) × Below 200% FPL” and the coefficient for “Active Violation (Month 1-6) × Below 200% FPL.” For a full-sample version of this table, see Appendix Table 34.

Appendix A - Learning About Subjective Uncertainty: Overinference from Observable Characteristics in Disaggregated Data

A.1 Experiment Priming Conditions

Each session was randomly assigned one of three priming conditions: a questionnaire on gender differences based on real-world survey data (*GenderDiffs*), a button-pressing task with group-based incentives (*ButtonTask*), or nothing as a control condition (*NoStage1*). In *NoStage1* sessions, subsequent stages of the experiment were relabeled so that participants were not aware other sessions had an additional stage. Across all sessions, 88 participants completed *GenderDiffs*, 64 participants completed *ButtonTask*, and 80 participants were assigned *NoStage1*.

In *GenderDiffs*, participants are asked to guess the percentage of men and women, respectively, to which a particular statement applies. Each statement is related to a health or economic outcome, such as life expectancy and average salary for workers in a particular industry, in which men and women significantly differ. The correct answers are drawn from the 2018 General Social Survey, U.S. Census Bureau, and CDC. Immediately after making their guesses, participants are shown the correct answers. At the end of the study, a bonus payment of $2 based on the binarized scoring rule (Hossain and Okui 2013) is calculated based on a randomly selected guess (i.e., one part, men or women, of one statement).

In *ButtonTask*, participants complete two five-minute rounds of the key-pressing task in Ariely et al. (2009) and DellaVigna and Pope (2018). The task requires participants to alternately press the “a” and “b” keys on their keyboard to earn payments. The incentive structure is inspired by Babcock et al. (2015). Specifically, in round 1 of the task, participants earn a piece rate per 100 key presses up to a threshold, and a one-time bonus for meeting or exceeding the threshold as an individual. In round 2, participants earn the same piece rate up to the threshold, but the bonus is only earned if *everyone in their group* meet or exceed the threshold. Participants are not informed of whether or not they earned this bonus until the end of the experiment session.[1]

GenderDiffs and *ButtonTask* are designed to prime different types of group identification. *GenderDiffs* calls participants' attention to the differences between men and women in various contexts. Because participants are shown the correct answers immediately after each guess, they receive 10 consecutive informative signals with different information about men and women; thus when they receive information about men and women in the decision stage, they might be habituated to discard information about the other gender as irrelevant to them. By contrast, *ButtonTask* encourages individuals to feel attached to their group in a way that is not explicitly related to gender. Instead of encouraging participants to disregard information about other genders, *ButtonTask* encourages participants to pay special attention to information about their group through a sense of camaraderie.

In summary, the randomized priming conditions are intended to shed light on the underlying mechanism of the result found in the decision stage. If participants in *GenderDiffs* sessions responded more strongly to in-group information than others, this would suggest identification with one's own gender (and lack of identification with other genders) drives the effect. Alternatively, if participants in *ButtonTask* responded more strongly, this would suggest identification with a particular assigned group, possibly irrespective of the trait on which that group assignment was based, drives the effect. Finally, if there is no difference between either condition and *NoStage1*, this implies that either baseline group identification is strong enough that priming is not required (i.e., a ceiling effect), or selective attention to in-group information is driven by a belief or trait which is not influenced by these priming tasks.

Tables 17 adds interaction terms between observed in-group and out-group wins, respec-

[1]The incentives paid in *ButtonTask* varied across the two sessions to which it was assigned. In the first, participants earned \$0.10 per 100 presses up to 1,000, and a \$1 bonus for meeting or exceeding 1,000 presses. In the second, because 100% of participants met or exceeded the threshold of 1,000 in the first session, incentives were reduced to \$0.05 per 100 presses up to 2,000 and a bonus of \$1 for meeting or exceeding 2,000 presses. While this means any analysis based on participants' actual performance in *ButtonTask* cannot be pooled across sessions, it should not significantly impact the priming effect of the stage.

tively, and the randomized priming conditions, the amount allocated to the in-group receiver, and self-reported attachment to one's own group throughout the experiment. While the interaction terms between observed in-group wins and assignment to *GenderDiffs* and *ButtonTask* respectively (with the *NoStage1* control condition as the reference group) have the expected positive sign, they are not statistically significant. This is despite the fact that the priming appeared to be effective: when asked what the experiment was about in the exit survey, *GenderDiffs* participants were 16 percentage points more likely to mention gender, while *ButtonTask* participants were 20.5 percentage points less likely compared to *NoStage1* (both differences $p < 0.05$). There are multiple possible interpretations of this result. First, the effect of observed in-group wins on an individual's X guess was quite strong in *NoStage1*; the non-interaction term in Table 17 implies that even without a priming task, observing an additional in-group win increased participants' average X guess by about 8.7 percentage points. Thus the lack of significance of priming could reflect a ceiling effect. Alternatively, taken at face value, the lack of a significant effect of priming may indicate that the types of group identification primed by the priming conditions are unrelated to the mechanism(s) behind selective attention to in-group information.

A.2 Observed Belief Updating and Bayesian Posteriors Continued

A.2.1 Explanation of Bayesian Posterior Calculation

To calculate Bayesian posteriors from participants' reported priors, it is necessary to estimate the respective probabilities of observing a signal within each interval conditional on the correct answer being within a particular interval, i.e., $P(\text{signal in interval } x \mid \text{correct answer in interval } y)\ \forall x, y \in \{1, 2, 3\}$. Additionally, because survey participants were asked to guess the *population* prevalence of a health condition among all individuals in a particular demographic group based on a nationally representative survey sample, it is necessary to take the survey sampling weights into account when estimating these probabilities, since the correct answer is the weighted proportion of individuals reporting a particular diagnosis in the sample.[2]

To estimate the necessary probabilities to calculate Bayesian posteriors, I implement a bootstrap approach with constructed counterfactual datasets derived from the NHIS which simulate the correct answer being in each respective interval. In the results reported in Section 1.5.1, I assume that participants' beliefs about the proportional likelihood of each integer within a particular interval are uniformly distributed; in other words, if a participant assigns probability x to the correct answer being in the first interval (0% to 8%), I assume they assign probability $\frac{x}{9}$ to each integer $\{0, 1, 2, ..., 7, 8\}$. In accordance with this, the counterfactual NHIS datasets simulate true prevalences that are uniformly distributed within each interval respectively. Later in this section, I report results based on an alternative assumption that participants' beliefs are skewed toward the closest values within each interval to the midpoint (12.5%) of the entire range of possible percentages (0% to 25%), thus using counterfactual datasets simulating true prevalences of 8%, 12.5%, and 17%, respectively.

These counterfactual datasets are constructed by replacing the value of the binary outcome variable for a randomly selected subsample of individuals within the relevant demo-

[2] If survey weights were not used, the expected value for the prevalence among a subsample of 50 would be the raw frequency of the condition in the survey sample rather than the population estimate.

graphic group (women, white Americans, college graduates, and 18-29 year olds) until the prevalence is contained within the intended interval. In other words, I first specify a target interval (low, middle, or high), then randomly select a specific target prevalence from the uniform distribution over that interval, and then randomly select a subset of individuals whose diagnoses (or lack thereof) are replaced in the counterfactual dataset so that the resulting prevalence approximates the target. As an example, in the true 2019 NHIS data, the prevalence of any type of cancer among women is approximately 10.2%. To construct a counterfactual dataset which simulates a prevalence in the low interval, I first make one draw of a target prevalence from the uniform distribution $U(0, 8.5)$.[3] Suppose the draw is 4%. I then randomly select approximately $\frac{10.2-4}{10.2} = 60.8\%$ of women in the NHIS who indicate a cancer diagnosis and replace this value with a non-diagnosis so that the resulting prevalence is approximately 4%. To simulate a higher prevalence, I replace values of the outcome variable for a random subset of individuals in the opposite way; for example, to simulate a prevalence of 21%, I randomly select approximately $\frac{10.8}{89.2} = 12.1\%$ of women *without* a cancer diagnosis and replace this value with a diagnosis. I then draw a random sample of 50 women from each of these counterfactual datasets and record the prevalence derived from this subsample. This entire process, including the random draw of target prevalence from the uniform distribution over the interval, random reassignment of diagnosis values, and the subsequent random subsample of 50, is repeated 10,000 times for each interval.

Under the assumption that participants correctly perceived the way the information was generated, the main effect of the information for a Bayesian updater is reducing the likelihood that the correct answer is contained in the diametrically opposed interval. In other words, receiving a signal in the low interval means it is less likely that the correct answer is in the

[3] I include prevalences up to, but not including, 8.5 because the correct answer is rounded to the nearest integer, meaning that e.g. a prevalence of 8.3% is contained in the low interval.

high interval, and vice-versa.[4] Most importantly, since the health condition and demographic group pairs in the survey were deliberately selected to be cases in which the true prevalence is nearly identical across groups, they should proportionally update their beliefs about *both* groups in the question if they have well-calibrated beliefs about group differences.

After calculating the likelihood of observing each possible signal, the Bayesian posterior for each interval is simply:

$$P\left(\text{correct answer in } x \mid \text{signal in } y\right) = \frac{P\left(\text{signal in } y \mid \text{correct answer in } x\right) \times P\left(\text{correct answer in } x\right)}{\sum_{m \in \{1,2,3\}} \Big(P\left(\text{signal in } y \mid \text{correct answer in } m\right) \times P\left(\text{correct answer in } m\right) \Big)}$$

Because all three possible signals have a nonzero probability of occurring regardless of the interval containing the correct answer, this posterior is well-defined everywhere. However, in cases where the prior assigns probability zero to the particular interval containing the signal, the Bayesian posterior is always zero, even though participants' posteriors may be likely to assign nonzero probability (Basieva et al. 2017). Since comparing observed posteriors to the Bayesian posterior is a primary objective of Section 1.5.1 and this section, this introduces a concern: if a sufficient number of participants assigned probability zero to a signal they received but do not follow Bayes' rule, including individuals with a prior of zero in the interval containing the signal may make belief updating look non-Bayesian on average, even if updating from nonzero priors is approximately Bayesian.[5] To address this concern, Table 18 replicates the specifications in Table 10 while excluding participants whose priors assigned

[4] The likelihood of the minimum prevalence among 10 random samples of 50 drawn from the NHIS being in the low interval rounds to 100% if the correct answer is in the low interval, 95% if the correct answer is in the middle interval, and 24% is the correct answer is in the high interval. Likewise, the likelihood of the maximum prevalence among 10 random samples of 50 drawn from the NHIS being in the high interval rounds to 100% if the correct answer is in the high interval, 93% if the correct answer is in the middle interval, and 14% if the correct answer is in the low interval. As a result, a low-prevalence signal is unlikely if the true prevalence is high and vice-versa, but in both cases, the information does not substantially distinguish between the interval that contains it and the middle interval.

[5] For example, an individual who assigned probability zero to the interval containing the signal and assigned a probability of 50% to this interval in the posterior is considered to have overinferred by 50 percentage points, which is a large positive outlier.

zero probability to the interval containing the signal. Among this subsample, the non-Bayesian difference in updating across groups remains highly statistically significant ($p < 0.01$).

A.2.2 Alternative Specification

The results reported in Section 1.5.1 assume that participants correctly identified the information as non-representative. However, because individuals' beliefs about the representativeness and/or informativeness of the information were not explicitly elicited, this is uncertain. To account for the possibility that participants considered the information representative and updated their beliefs accordingly, the results of this subsection compare observed posteriors to Bayesian posteriors based on the assumption that the information was derived from a single random sample drawn from the NHIS of 50 individuals in a particular group.

Under the assumptions of uniform within-interval beliefs and participants perceiving the information as the result of one random draw from the NHIS, the respective probabilities of receiving a signal which coincides with the simulated correct answer ($P(\text{signal in interval } x \mid \text{correct answer in interval } x)\ \forall x \in \{1, 2, 3\}$) are approximately 81% for the low interval, 50% for the middle interval, and 71% for the high interval. As a result, if participants are Bayesian, they should increase the probability they assign to the interval containing the signal, especially since their priors are quite flat. Additionally, since the health condition and demographic group pairs in the survey were deliberately selected to be cases in which the true prevalence is nearly identical across groups, they should proportionally update their beliefs about *both* groups in the question if they have well-calibrated beliefs about group differences. The results in Tables 19 and 20 confirm that this key result holds under this alternative assumption, since the coefficient on "Group represented in information" remains highly statistically significant across all columns of both tables.

A.3 Additional Tables

Table 15: Gender Differences in Belief Updating and Switch Point Choices

	Believed X	Price List Switch Point
Observed in-group test run wins	10.123***	0.443**
	(1.502)	(0.214)
In-group wins × Male indicator	0.632	0.027
	(2.129)	(0.319)
Observed out-group test run wins	-0.811	0.212
	(1.776)	(0.192)
Out-group wins × Male indicator	1.830	-0.207
	(2.372)	(0.254)
Male indicator	1.507	0.732
	(4.786)	(0.568)
Age in years	-0.264	-0.006
	(0.361)	(0.063)
Observations	232	232
Adjusted R^2	0.244	0.032

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Group-clustered standard errors in parentheses. "Number of in-group (out-group) test run wins observed" refers to the three in-group (out-group) hypothetical draws of the lottery with X chance of winning $5 (based on others' unobserved assigned X values) which participants observed prior to guessing their own X. Participants were informed that these draws were for informational purposes only and did not affect others' payoffs. "Reported belief about own X" was reported on a slider over integers between 0 to 100. "In-group information on left" is an indicator variable which equals 1 when the participant saw the information about their own group in the left-hand column of the table in which test run results were reported. For an example of this table, see Figure 1. All regressions include controls for priming condition. For an explanation of the priming conditions, see Appendix Section A.1.

Table 16: Alternative Session-Level Belief Regressions

	Own group's X was higher	Own group's X was lower
In-group test run wins - Out-group test run wins	0.155***	-0.207***
	(0.029)	(0.024)
(In-group test run wins - Out-group test run wins) × Male indicator	0.008	0.115***
	(0.038)	(0.035)
Male indicator	0.173***	-0.283***
	(0.049)	(0.052)
Observations	231	231
Adjusted R^2	0.219	0.273

Notes: * p < 0.1, ** p < 0.05, *** p < 0.01. Group-clustered standard errors in parentheses. "In-group test run wins - Out-group test run wins" refers to the difference in outcomes (an integer between -3 and 3) between the three in-group and three out-group outcomes among the six hypo-thetical draws of the lottery with X chance of winning $5 (based on others' unobserved assigned X values) which participants observed prior to guessing their own X. For example, if the participant observed 2 wins among their own group and 1 win among the other group, this variable takes a value of 2 - 1 = 1. Participants were informed that these draws were for informational purposes only and did not affect others' payoffs. The "Belief about average X for men and women" was reported on a 5-point Likert scale from "Women's X was much higher than men's" to "Men's X was much higher than women's" and then assigned an integer value between 1 and 5. To form the dependent variables in these regressions, two indicator variables were created: one for participants believing their own gender's assigned X was at least slightly higher than the other gender's on average (i.e., women answering 1 or 2, and men answering 4 or 5), and one for the opposite (i.e., women answering 4 or 5, and men answering 1 or 2). Note that these regressions have one fewer observation (231 vs 232) than previous tables because one participant disconnected from the session before answering this question. All regressions included controls for the priming condition. For explanations of the priming conditions, see Appendix Section A.1.

Table 17: Effect of Priming Conditions on Responses to Test Run Wins

	(1) Believed X
Observed in-group test run wins	8.721***
	(1.280)
In-group wins × Gender differences questionnaire	2.413
	(2.374)
In-group wins × Button-pressing task	3.477
	(2.651)
Observed out-group test run wins	0.046
	(1.915)
Out-group wins × Gender differences questionnaire	1.006
	(2.644)
Out-group wins × Button-pressing task	-1.373
	(3.029)
Priming: Gender differences questionnaire	-5.892
	(5.788)
Priming: Button-pressing task	-2.972
	(5.407)
Observations	232
Adjusted R^2	0.242

Notes: * p < 0.1, ** p < 0.05, *** p < 0.01. Group-clustered standard errors in parentheses. "Number of in-group (out-group) test run wins observed" refers to the three in-group (out-group) hypothetical draws of the lottery with X chance of winning \$5 (based on others' unobserved assigned X values) which participants observed prior to guessing their own X. Participants were informed that these draws were for informational purposes only and did not affect others' payoffs. "Reported belief about own X" was reported on a slider over integers between 0 to 100. For an explanation of each priming condition, see Section 1.2. Controls for gender and age are included.

Table 18: Comparing Observed Posteriors to Bayesian Predictions: Baseline Model (Excludes Zero Priors)

Dependent variable: Difference (error) in probability assigned to interval containing the information between observed posterior and Bayesian prediction				
Constant	-0.033*** (0.009)	-0.039*** (0.014)	-0.003 (0.015)	-0.003 (0.025)
Group in info	0.114*** (0.013)	0.080*** (0.019)	0.210*** (0.021)	0.188*** (0.035)
Respondent is in group in info		0.011 (0.019)		0.000 (0.031)
Group in info X Participant is in group in info		0.064** (0.026)		0.036 (0.044)
Sample	All	All	Updaters	Updaters
Num.Obs	1493	1493	747	747
R2 Adj.	0.047	0.056	0.118	0.117

Notes: * p < 0.1, ** p < 0.05, *** p < 0.01. The constant represents the average error relative to the Bayesian predicted posterior in the group not represented in the information. Negative values indicate under-inference while positive values indicate over-inference. "Group in info" is an indicator variable which equals 1 when the observation contains a belief about the group to which the information pertained. For example, in the cancer prevalence by gender questiion, "Group in info" equals 1 for beliefs about women and 0 for beliefs about men. "Participant is in group represented in information" is an indicator variable which equals 1 when the *participant* to which the observation pertains is in the group to which the information pertained, regardless of which group the belief is about. In the same cancer by gender example, "Participant is in group in info" equals 1 for all female participants' beliefs about men *and* women. For an explanation of how these Bayesian predictions were calculated, see Appendix Section A.2.1.

Table 19: Comparing Observed Posteriors to Bayesian Predictions: Alternative Model

Dependent variable: Difference (error) in probability assigned to interval containing the information between observed posterior and Bayesian prediction				
Constant	-0.291*** (0.011)	-0.283*** (0.017)	-0.247*** (0.018)	-0.215*** (0.030)
Group in info	0.120*** (0.016)	0.060** (0.024)	0.234*** (0.026)	0.180*** (0.043)
Participant is in group in info		-0.015 (0.023)		-0.052 (0.038)
Group in info X Participant is in group in info		0.112*** (0.032)		0.087 (0.054)
Sample	All	All	Updaters	Updaters
Num.Obs	1616	1616	834	834
R2 Adj.	0.033	0.043	0.088	0.088

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. The constant represents the average error relative to the Bayesian predicted posterior in the group not represented in the information. Negative values indicate under-inference while positive values indicate over-inference. "Group in info" is an indicator variable which equals 1 when the observation contains a belief about the group to which the information pertained. For example, in the cancer prevalence by gender questiion, "Group in info" equals 1 for beliefs about women and 0 for beliefs about men. "Participant is in group represented in information" is an indicator variable which equals 1 when the *participant* to which the observation pertains is in the group to which the information pertained, regardless of which group the belief is about. In the same cancer by gender example, "Participant is in group in info" equals 1 for all female participants' beliefs about men *and* women. For an explanation of how these Bayesian predictions were calculated, see Appendix Section A.2.1.

Table 20: Comparing Observed Posteriors to Bayesian Predictions: Alternative Model (Excludes Zero Priors)

Dependent variable: Difference (error) in probability assigned to interval containing the information between observed posterior and Bayesian prediction				
Constant	-0.336*** (0.010)	-0.331*** (0.014)	-0.311*** (0.016)	-0.296*** (0.026)
Group in info	0.112*** (0.014)	0.061*** (0.020)	0.216*** (0.022)	0.177*** (0.037)
Respondent is in group in info		-0.009 (0.019)		-0.023 (0.033)
Group in info X Participant is in group in info		0.096*** (0.027)		0.063 (0.046)
Sample	All	All	Updaters	Updaters
Num.Obs	1493	1493	747	747
R2 Adj.	0.042	0.054	0.112	0.112

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. The constant represents the average error relative to the Bayesian predicted posterior in the group not represented in the information. Negative values indicate under-inference while positive values indicate over-inference. "Group in info" is an indicator variable which equals 1 when the observation contains a belief about the group to which the information pertained. For example, in the cancer prevalence by gender question, "Group in info" equals 1 for beliefs about women and 0 for beliefs about men. "Participant is in group represented in information" is an indicator variable which equals 1 when the *participant* to which the observation pertains is in the group to which the information pertained, regardless of which group the belief is about. In the same cancer by gender example, "Participant is in group in info" equals 1 for all female participants' beliefs about men *and* women. For an explanation of how these Bayesian predictions were calculated, see Appendix Section A.2.1.

A.4 Additional Figures

Figure 16: Main Specification Bayesian Comparisons, Cancer by Gender, All Participants

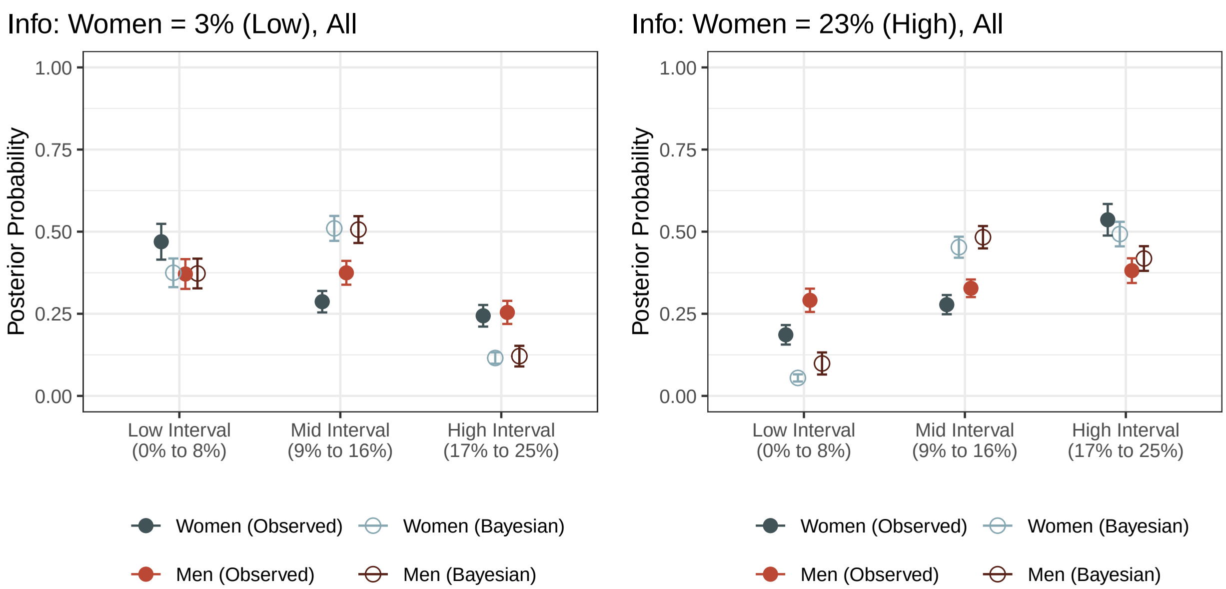


Figure 17: Main Specification Bayesian Comparisons, Diabetes by Race, All Participants

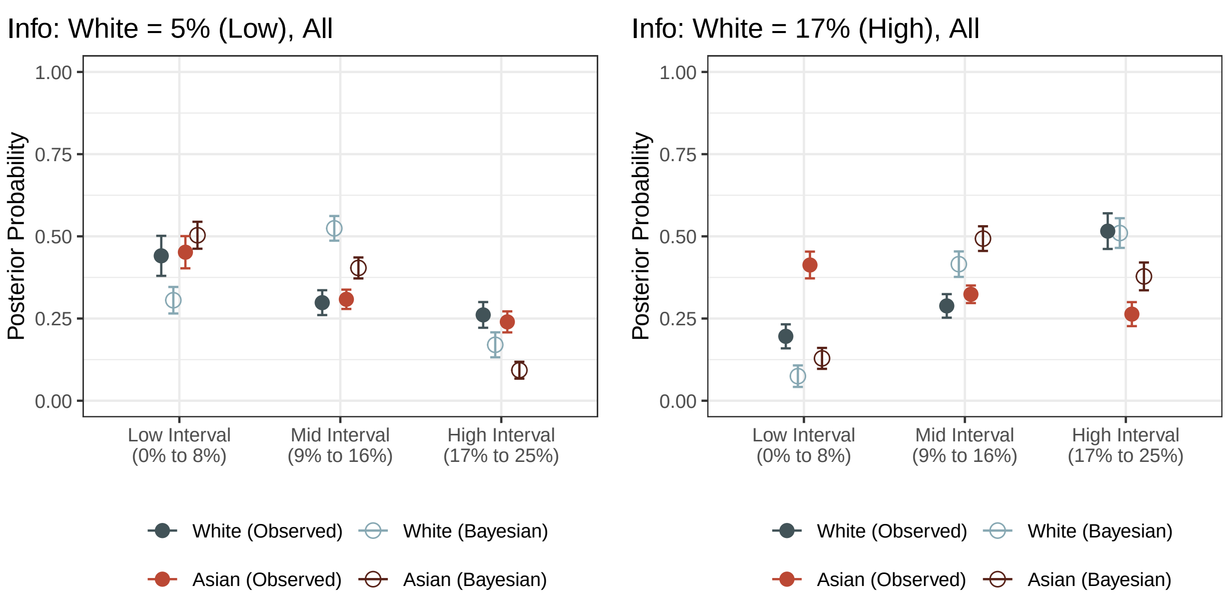


Figure 18: Main Specification Bayesian Comparisons, Anxiety by Education, All Participants

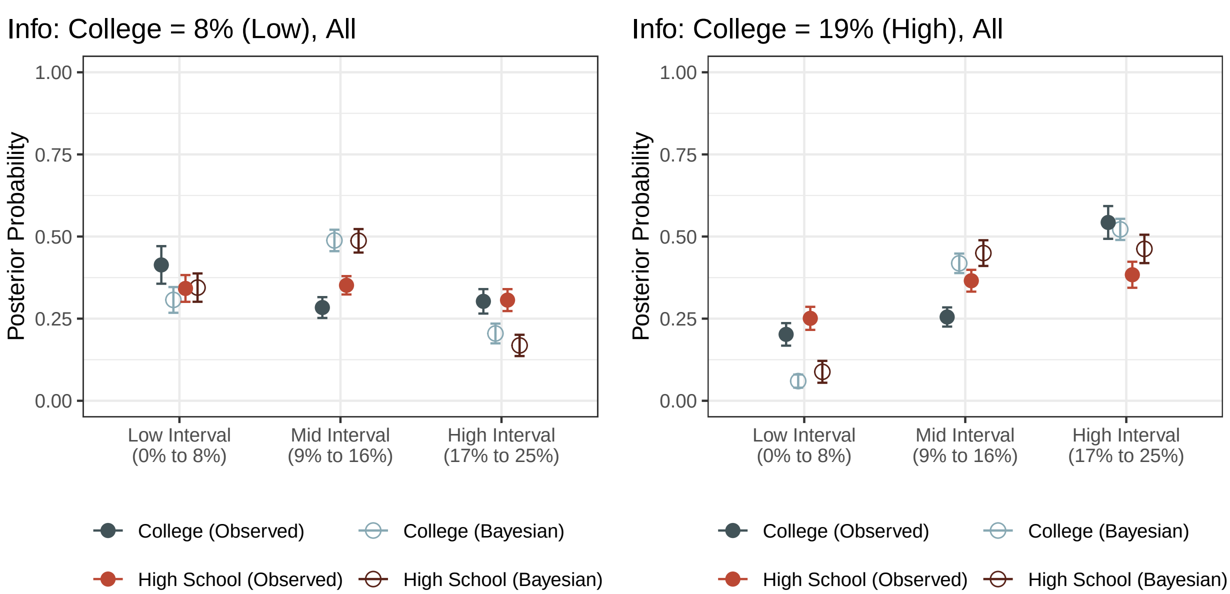


Figure 19: Main Specification Bayesian Comparisons, Depression by Age, All Participants

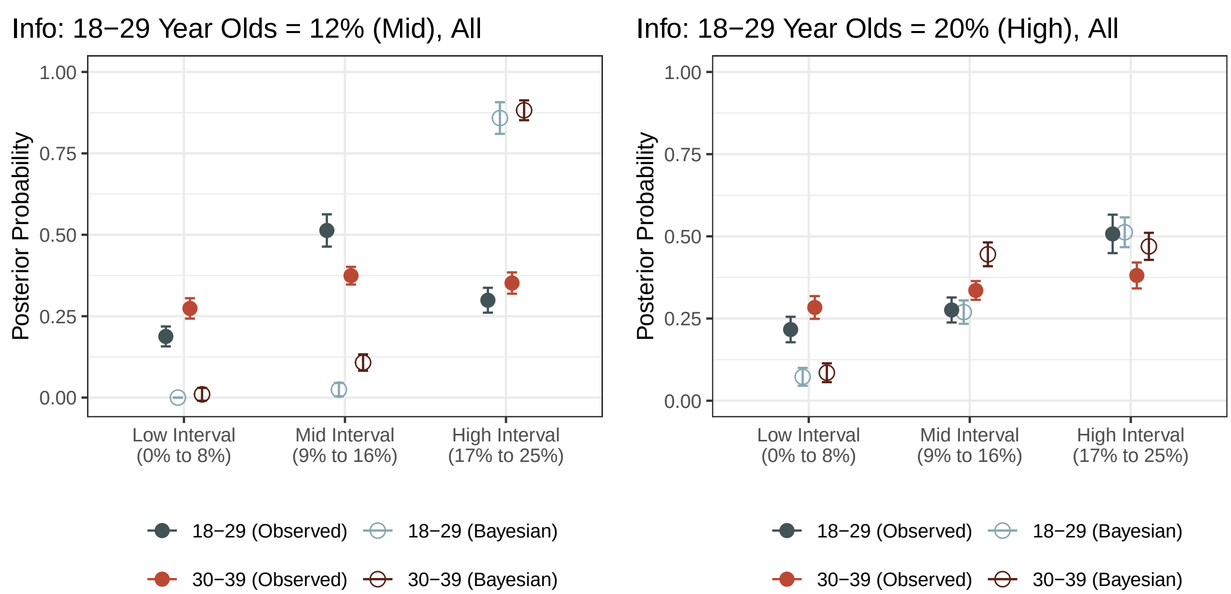


Figure 20: Main Specification Bayesian Comparisons, Cancer by Gender, Updaters Only

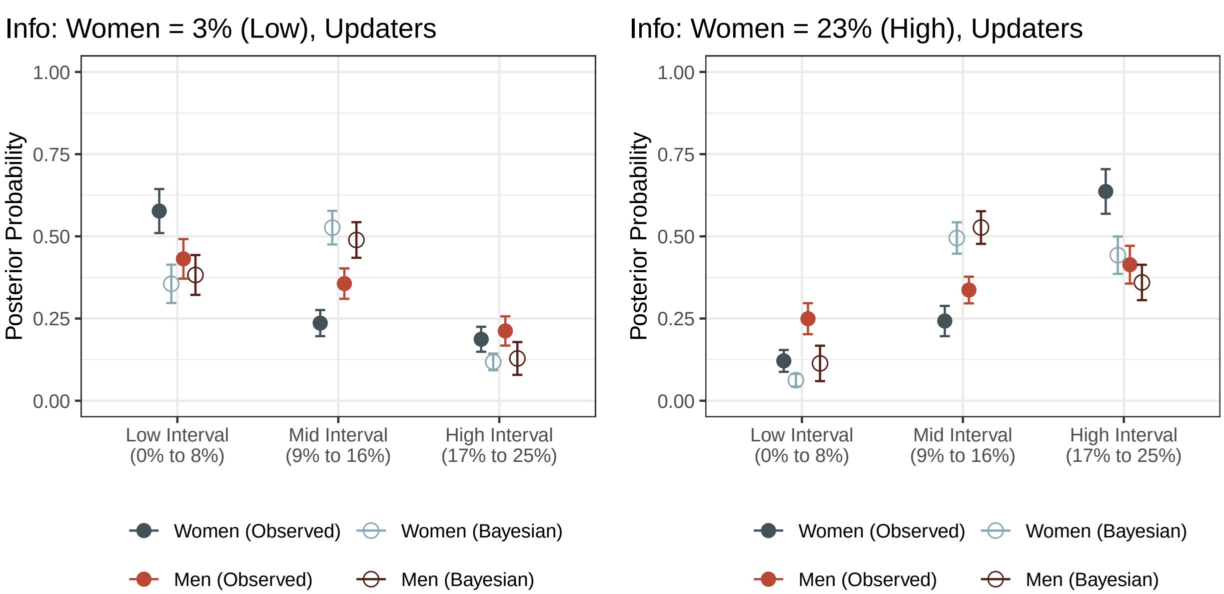


Figure 21: Main Specification Bayesian Comparisons, Diabetes by Race, Updaters Only

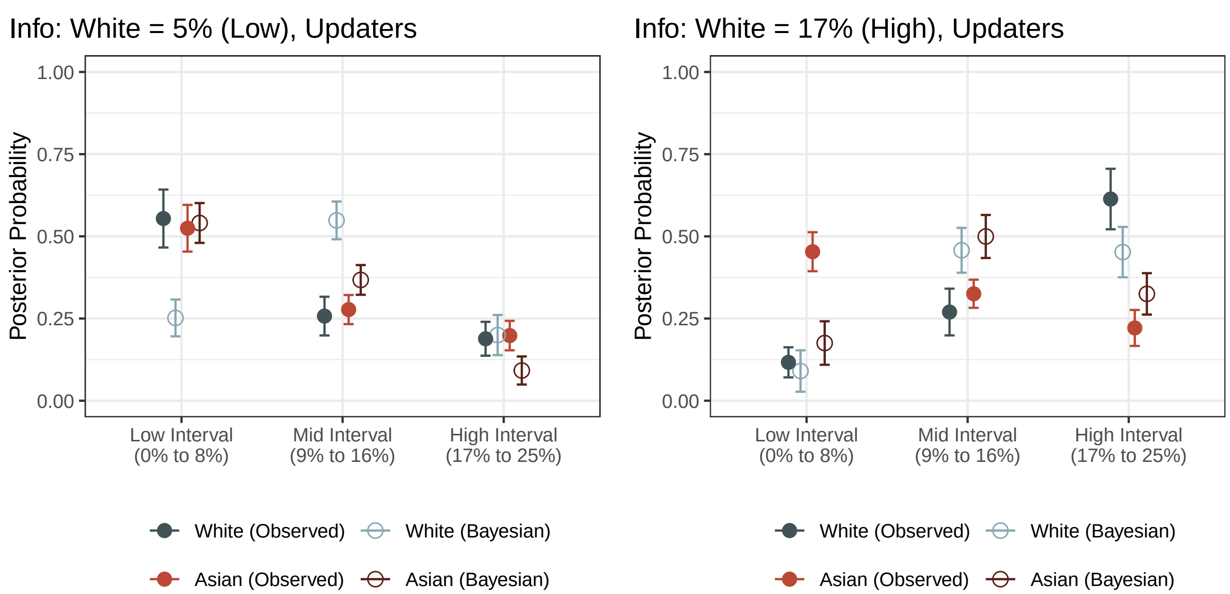


Figure 22: Main Specification Bayesian Comparisons, Anxiety by Education, Updaters Only

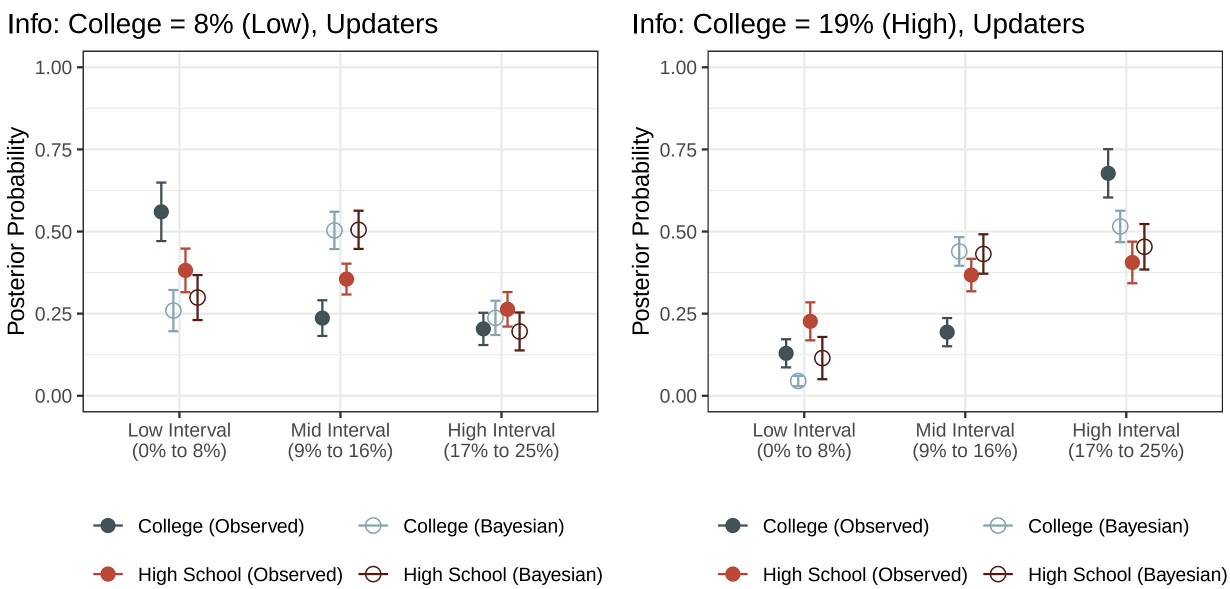


Figure 23: Main Specification Bayesian Comparisons, Depression by Age, Updaters Only

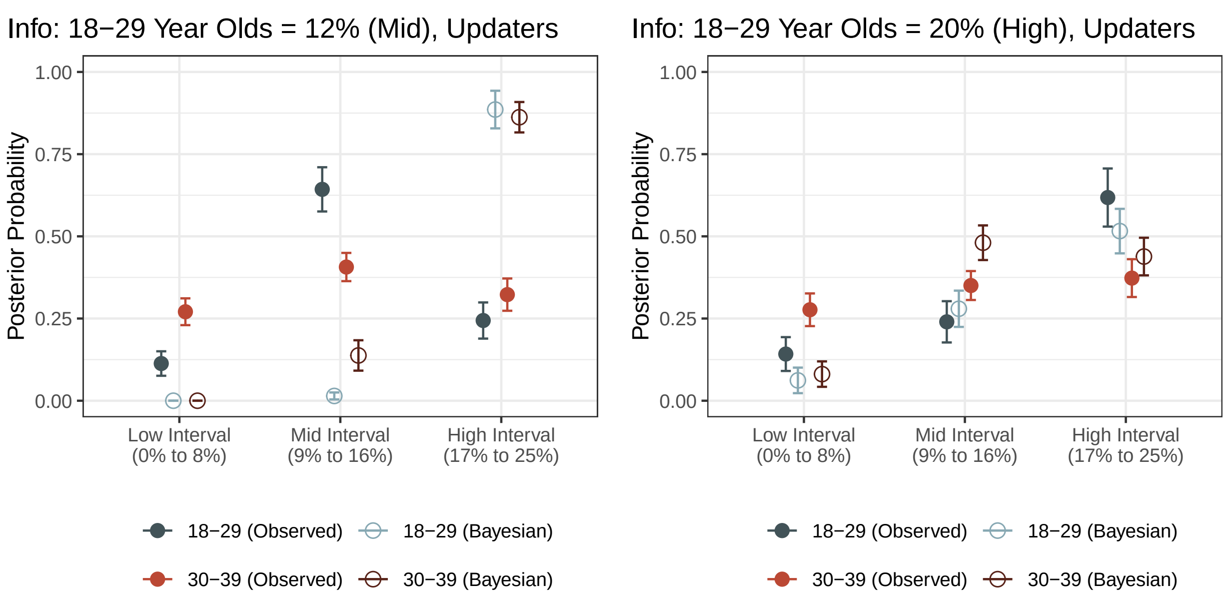


Figure 24: Alternative Specification Bayesian Comparisons, Diabetes by Race, All Participants

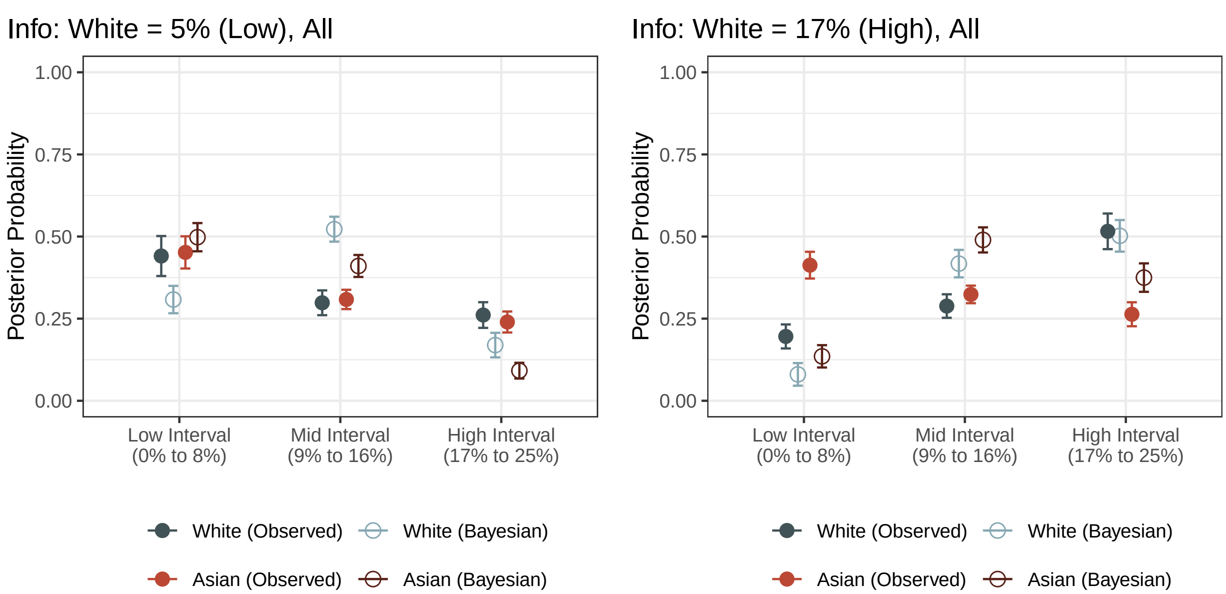


Figure 25: Alternative Specification Bayesian Comparisons, Anxiety by Education, All Participants

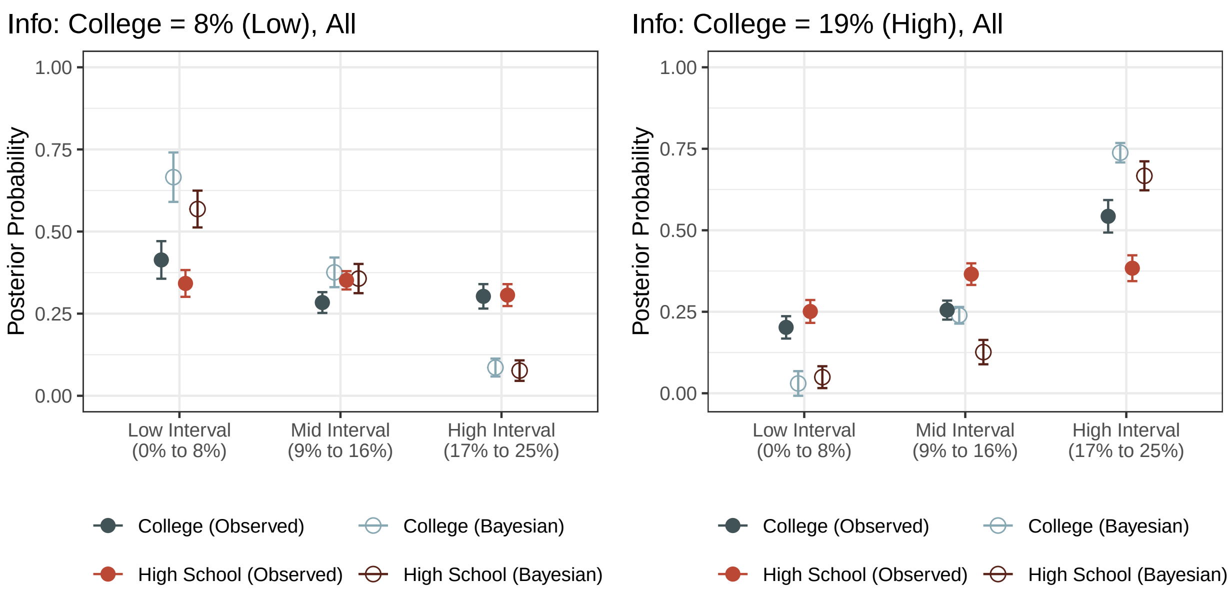


Figure 26: Alternative Specification Bayesian Comparisons, Depression by Age, All Participants

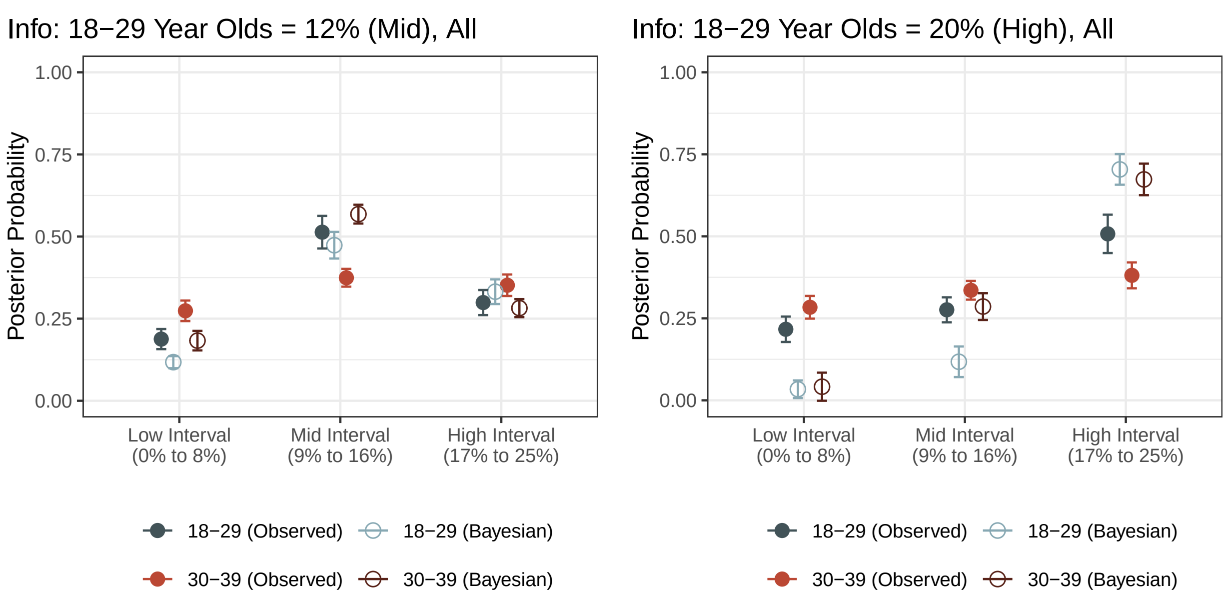


Figure 27: Alternative Specification Bayesian Comparisons, Cancer by Gender, Updaters Only

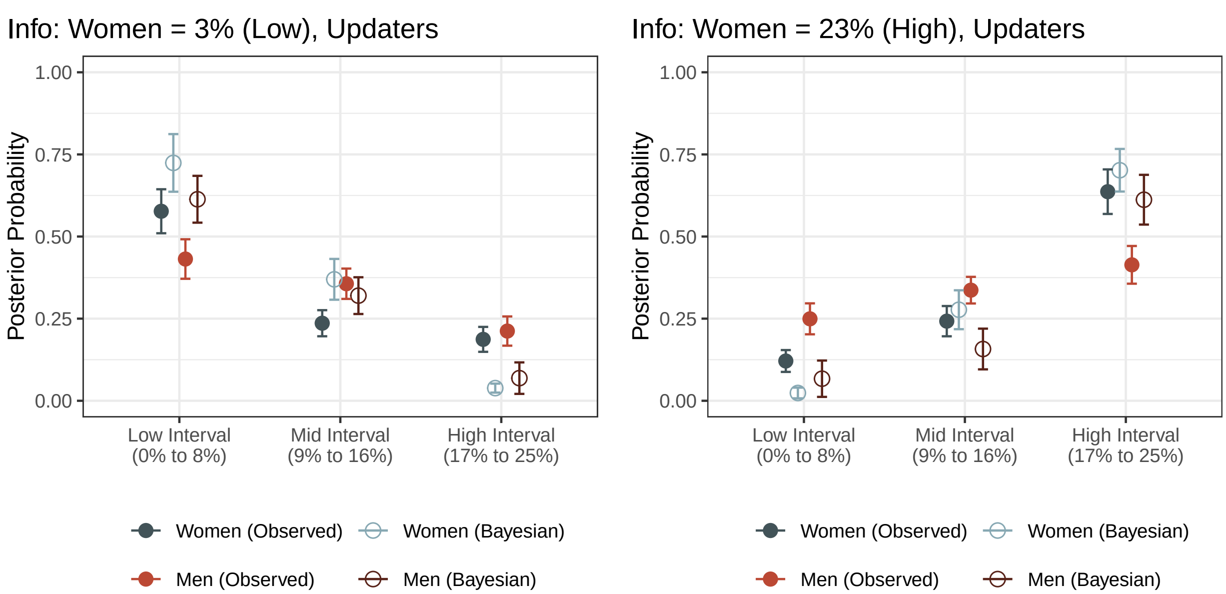


Figure 28: Alternative Specification Bayesian Comparisons, Diabetes by Race, Updaters Only

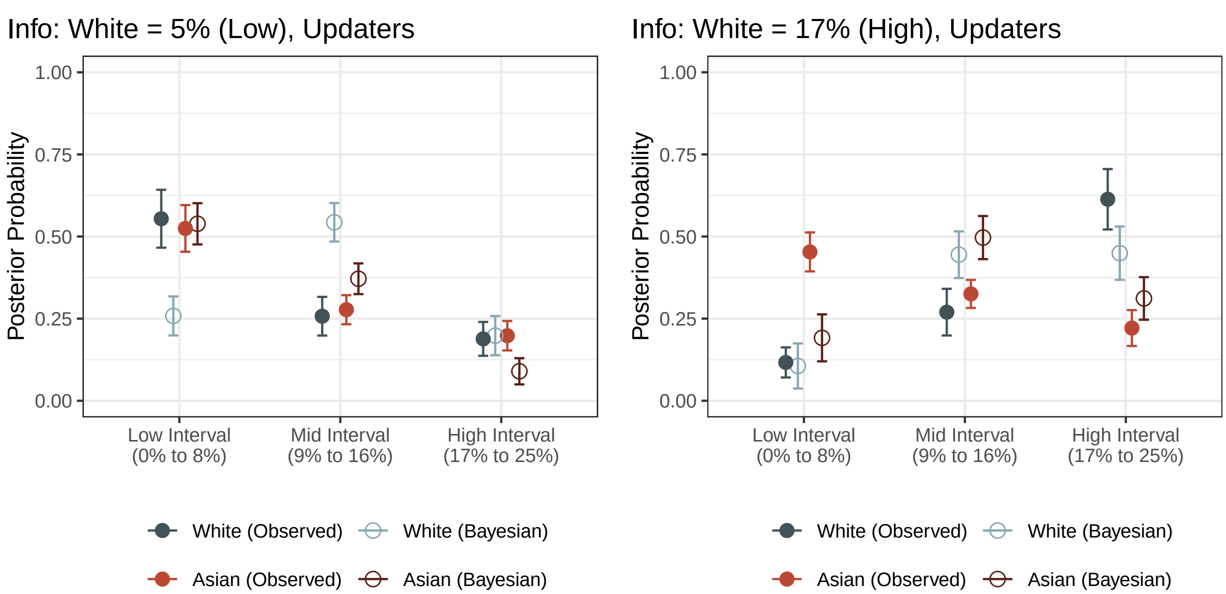


Figure 29: Alternative Specification Bayesian Comparisons, Anxiety by Education, Updaters Only

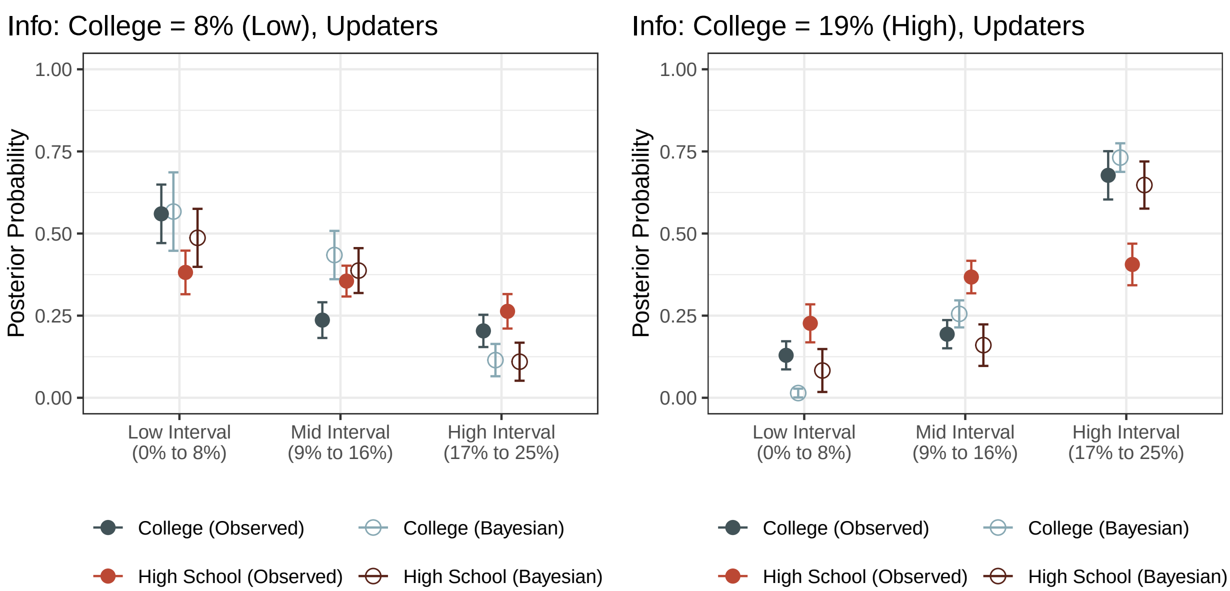


Figure 30: Alternative Specification Bayesian Comparisons, Depression by Age, Updaters Only

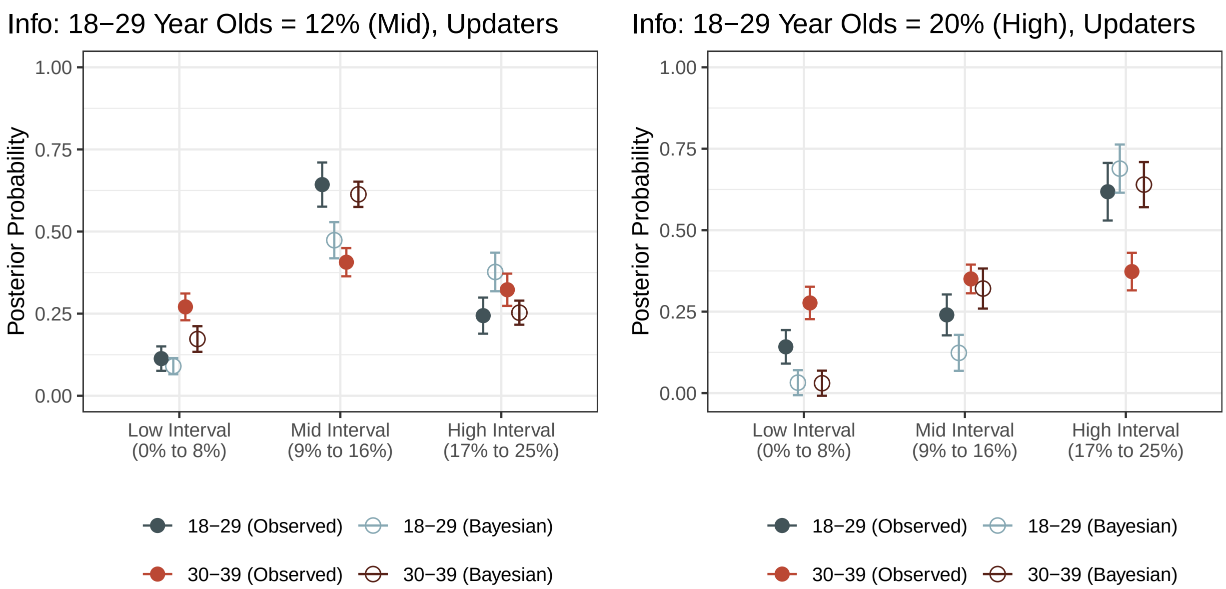


Figure 31: Survey Belief Updating Figures, Cancer and Diabetes (All Respondents)

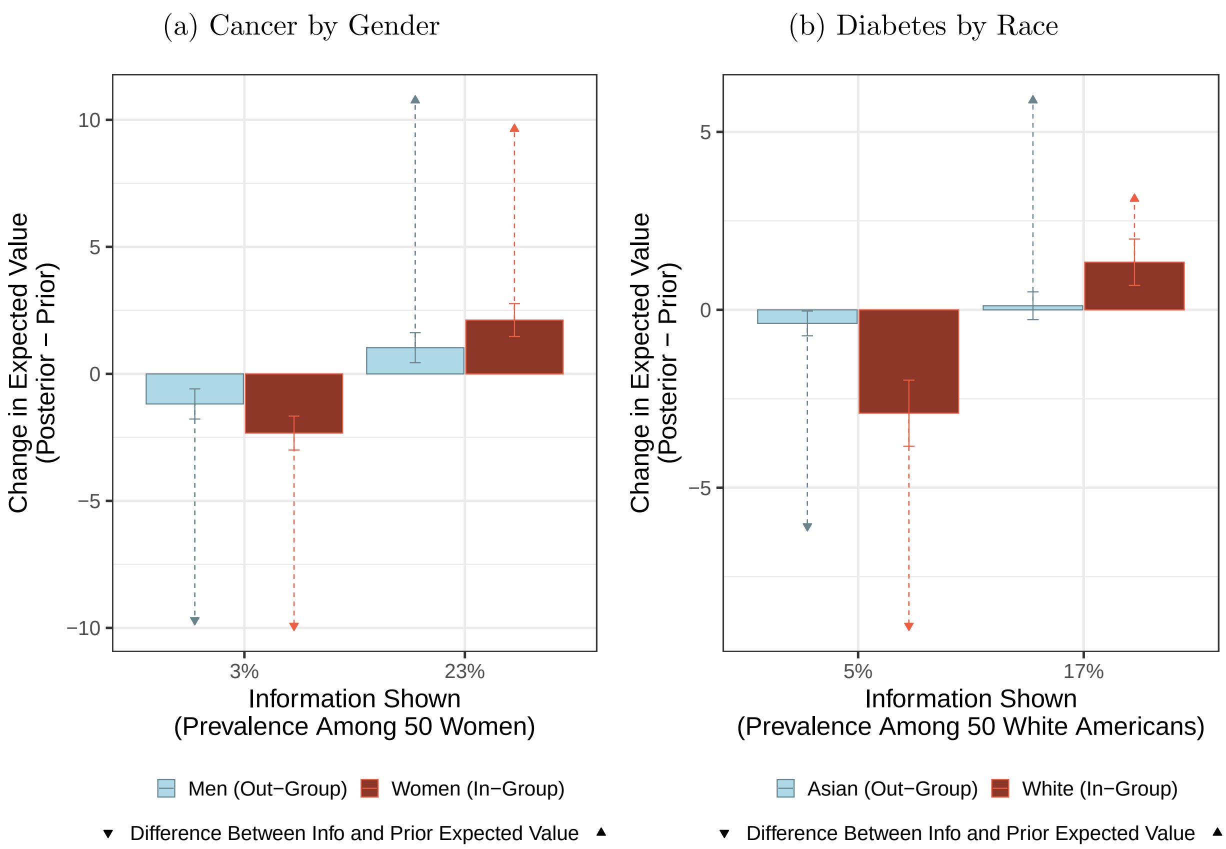
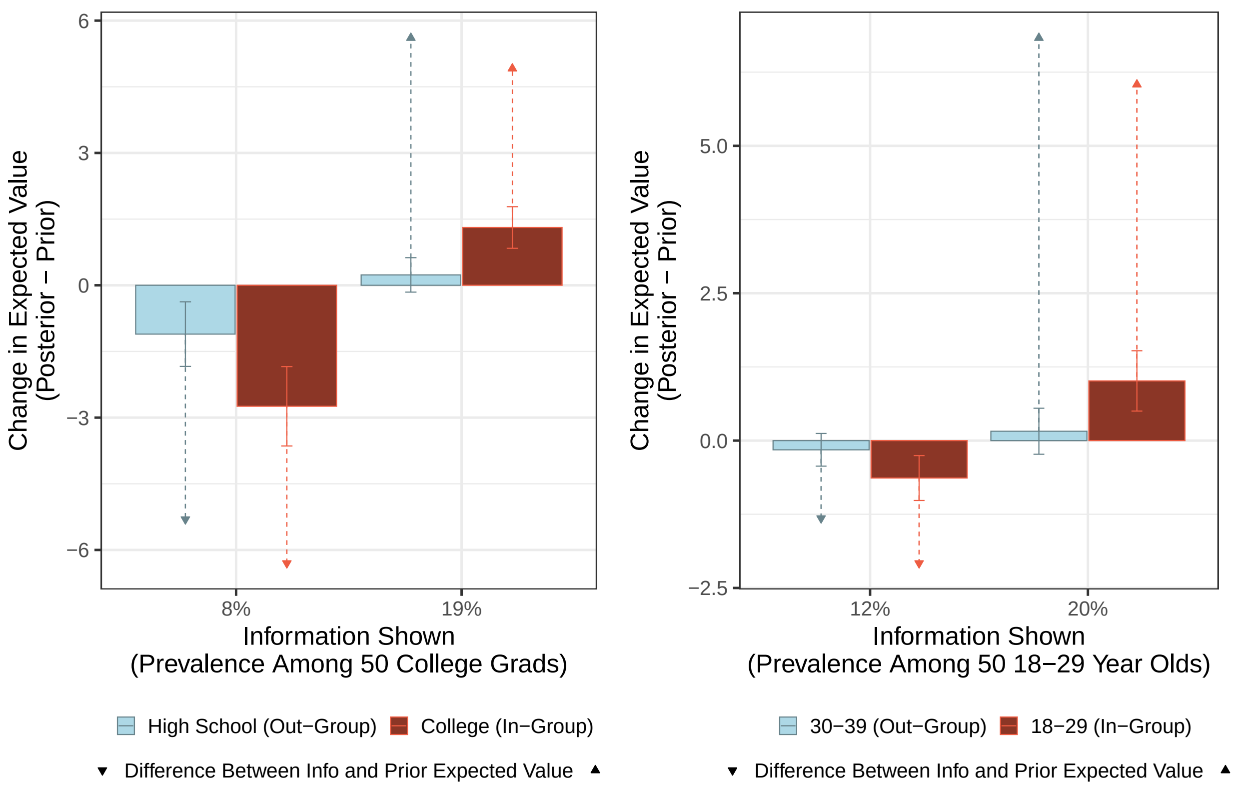


Figure 32: Survey Belief Updating Figures, Anxiety and Depression (All Respondents)

(a) Anxiety by Education

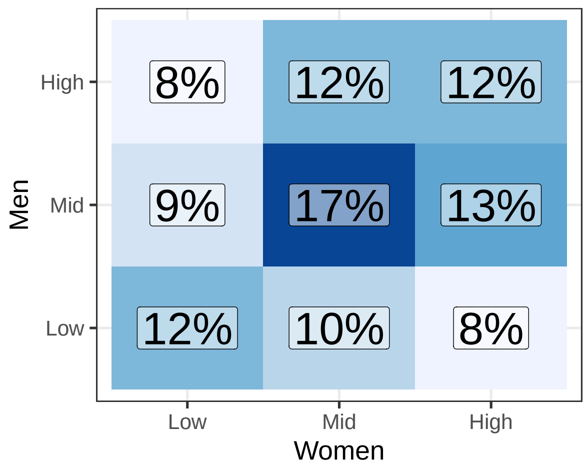
(b) Depression by Age Group



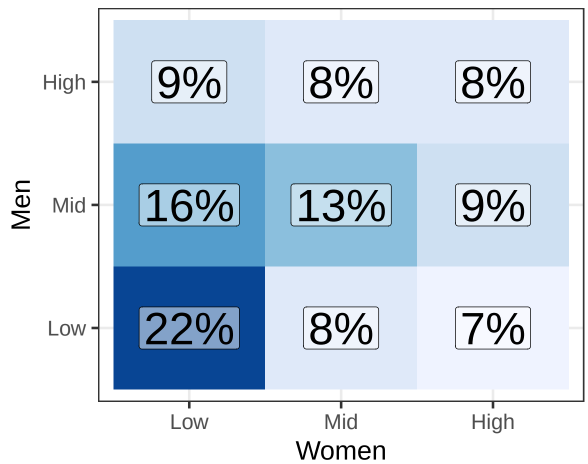
A.4.1 Survey Prior and Posterior Figures

Figure 33: Cancer Prevalence by Gender

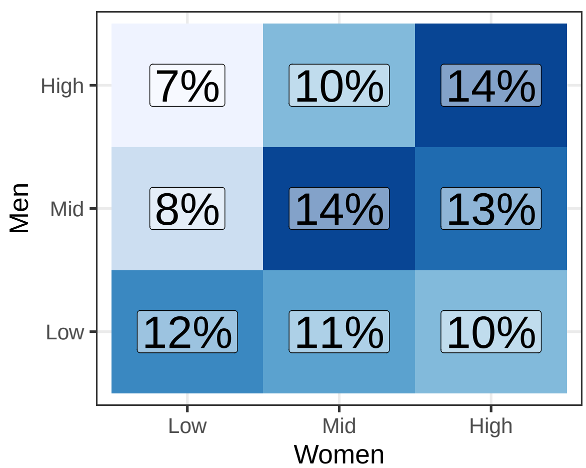
(a) Prior (Info: Women = 3%)



(b) Posterior (Info: Women = 3%)



(c) Prior (Info: Women = 23%)



(d) Posterior (Info: Women = 23%)

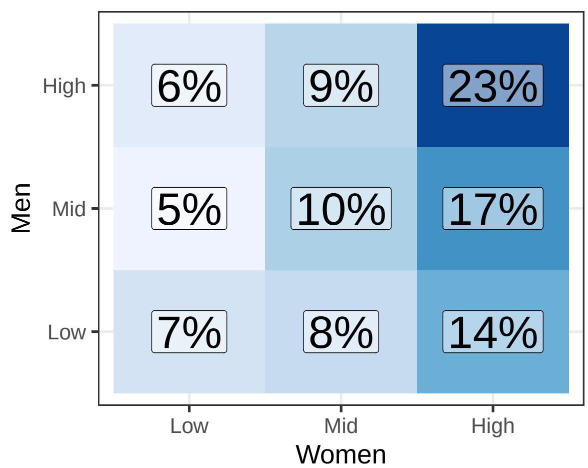
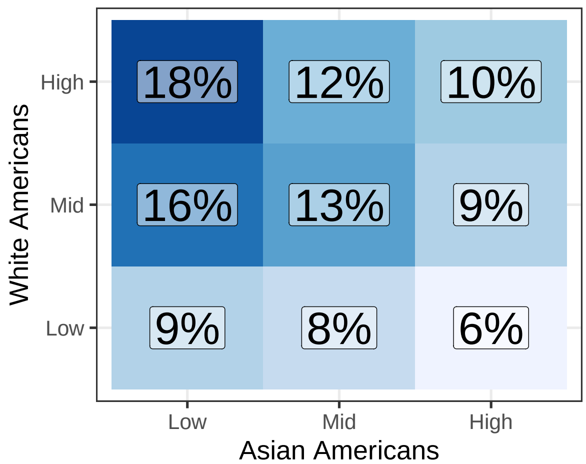
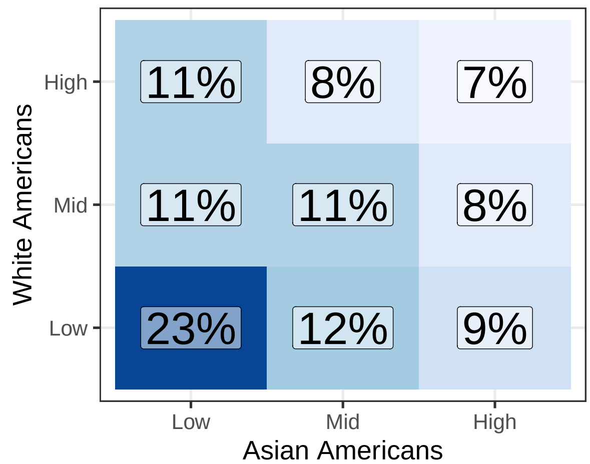


Figure 34: Diabetes Prevalence by Race

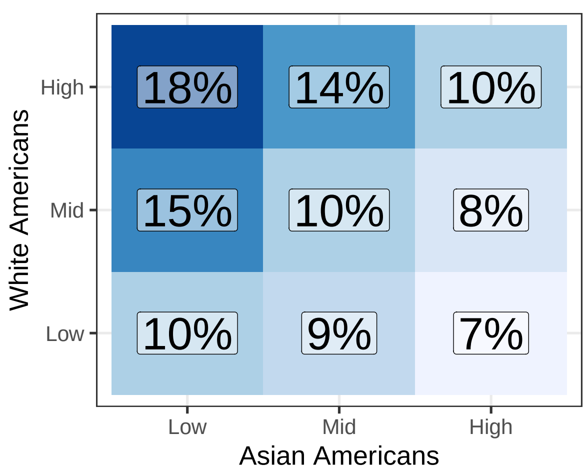
(a) Prior (Info: White = 5%)



(b) Posterior (Info: White = 5%)



(c) Prior (Info: White = 17%)



(d) Posterior (Info: White = 17%)

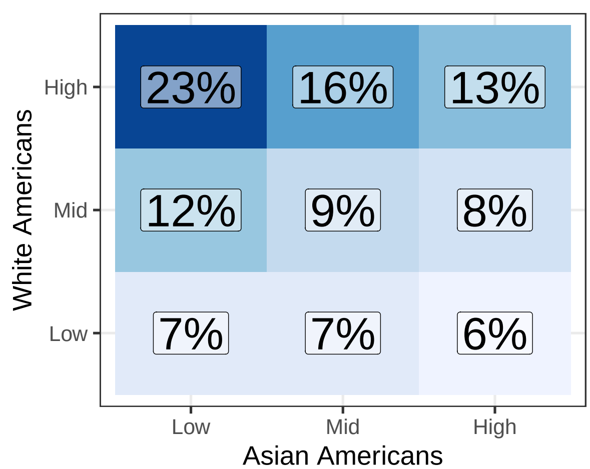
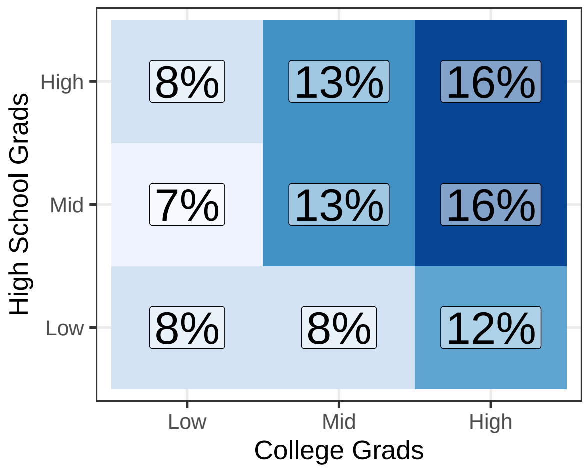
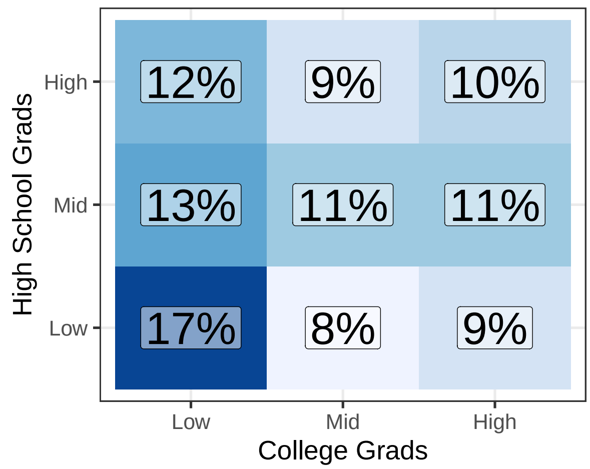


Figure 35: Anxiety Prevalence by Education

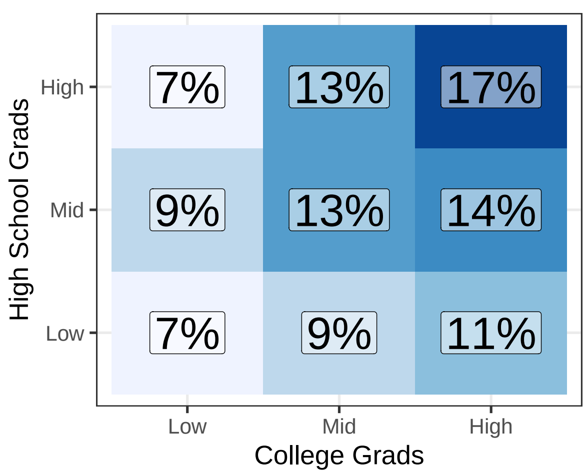
(a) Prior (Info: College = 8%)



(b) Posterior (Info: College = 8%)



(c) Prior (Info: College = 19%)



(d) Posterior (Info: College = 19%)

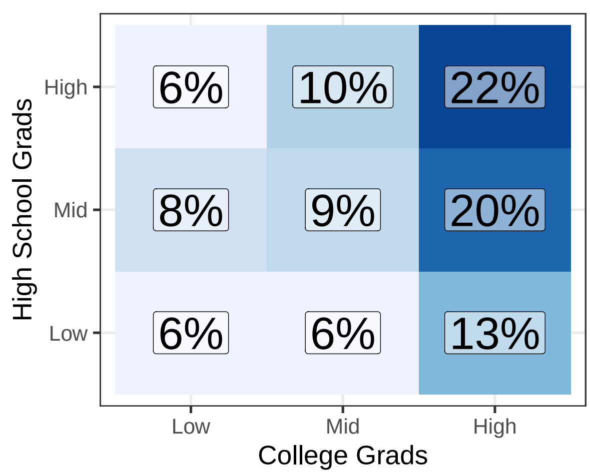
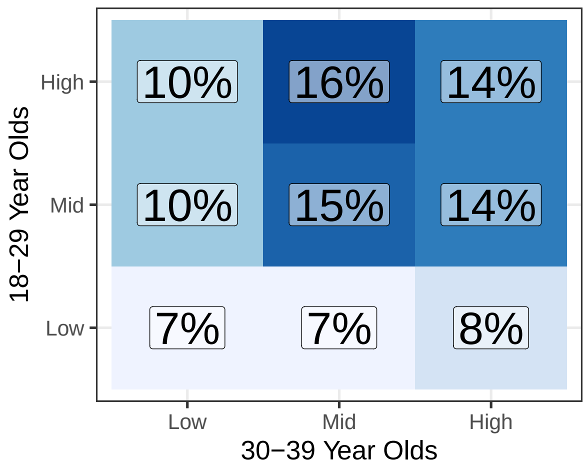
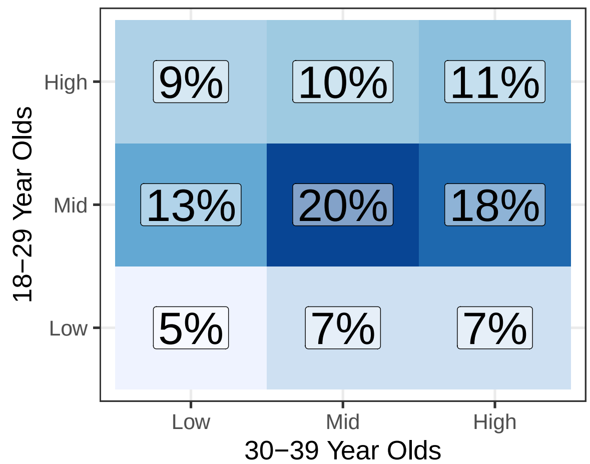


Figure 36: Depression Prevalence by Age Group

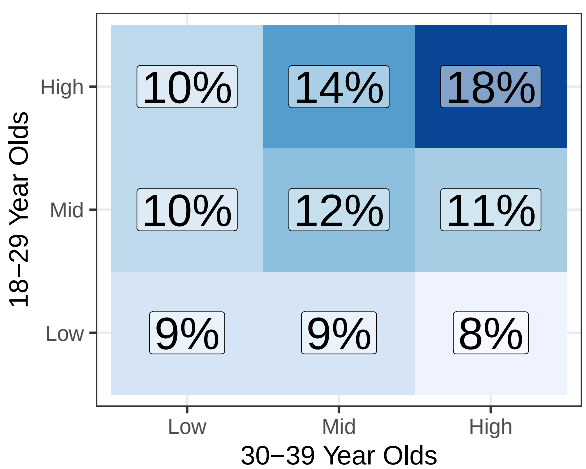
(a) Prior (Info: 18-29 Year Olds = 12%)



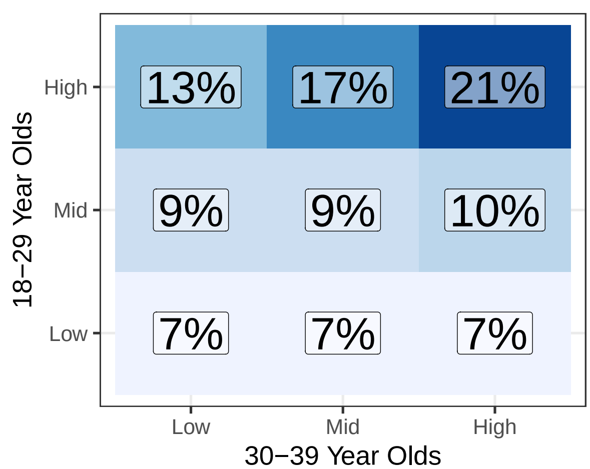
(b) Posterior (Info: 18-29 Year Olds = 12%)



(c) Prior (Info: 18-29 Year Olds = 20%)



(d) Posterior (Info: 18-29 Year Olds = 20%)



Appendix B - Water Availability and Heat-Related Mortality: Evidence from South Africa

B.1 Heterogeneity

I hypothesize two main types of heterogeneity in the effect of water availability on the heat-mortality relationship. First, in South Africa, water bearers are almost exclusively women and children (Graham et al. 2016). As natural water availability increases, households are more likely to have a sufficient water source nearby. Thus in households that rely on external sources, the physical cost of retrieving water decreases. Since heat exposure increases the risk of injury during physical exertion (Nelson et al. 2011), water availability is likely to have an especially strong effect on the heat-related mortality risk of water bearers.

Second, the degree to which water availability reduces the heat-related mortality risk of a household is likely influenced by its primary source of water, particularly whether that source is internal or external to the household. However, the direction of this heterogeneity is uncertain, and depends on the predominant mechanisms. Households with a piped water connection are less likely to be affected by variation in water availability on the intensive margin; the faucet either turns on or it does not. For households relying on external sources of water, in addition to decreasing the physical cost of retrieving water, increased natural water availability also decreases stagnation in smaller rivers. Since stagnant water has a higher risk of contamination, an increase in water availability is likely to decrease the risk of waterborne disease and resulting mortality. However, households with a piped water connection could be even more vulnerable to variation in water availability on the extensive margin, since they are unaccustomed to retrieving water from outside the home and may need to do so suddenly and unexpectedly if their pipes shut off. This source of heterogeneity may lead to heterogeneous estimated effects by population group, since as discussed in section 2, Black South Africans are less than half as likely to have piped water in-residence than their white counterparts, while Coloured[1] South Africans are less than three-quarters as likely.

Tables 21 through 24 test for the expected heterogeneous treatment effects. As expected, the mitigating effect of upstream water availability on the heat-mortality relationship is significantly larger for women. Because the dam level measures are mean-centered, the non-interaction coefficient on CDD 90°F is the estimated effect of one CDD 90°F at the status quo average of water availability; this coefficient is indistinguishable from zero for men, but positive and highly significant ($p < 0.01$) for women. This suggests that variation in water availability along the intensive margin is more significant for women's mortality risk during heat waves, which is consistent with their higher likelihood of being the primary water bearer for households relying on external sources and the associated hypothesis for water bearers.

The results in Table 22 on population group are less clear, but also appear broadly consistent with the intensive and extensive margin hypotheses for households with and without a potable water source on-premises, respectively. The non-interaction coefficient on CDD 90°F is positive and very close to significance at conventional levels ($p = 0.104$ and $p = 0.16$, respectively) for Black Africans and Coloured South Africans, while the coefficient of the interaction between upstream water availability and CDD 90°F is negative and significant ($p < 0.05$) for both. While the interaction coefficient is also negative and significant ($p < 0.01$) for White South Africans, the non-interaction coefficient on CDD 90°F is indistinguishable from zero. Similar to the results by sex, this suggests that variation in upstream water availability on the intensive margin is more significant to the mortality risk of Black African and Coloured South Africans than for White South Africans, which is consistent with their lower likelihood of having a piped water source on-premises and the associated hypothesis for households relying on external water sources. In other words, to eliminate increases in mortality risk during heat waves, maintaining the historical average is sufficient for White South Africans, while further increases above the historical average are necessary for Black African and Coloured South Africans.

Table 23 presents an especially strong mitigating effect on infant mortality, with a coef-

[1] the official term used by the South African government for individuals of mixed African and European descent

ficient size of about 28.6% of the mean infant mortality rate per million. Given the intuitive relationship between women's health outcomes and the health outcomes of infants, this is consistent with the previous finding of stronger effects for women. The mitigating effects are fairly proportional across the rest of the age distribution except for those above age 65, for whom the mitigating effect is (albeit barely) not significant at conventional levels ($p \sim 0.12$) and the point estimate is only about 2.4% of the mean. For this age group, the estimated mitigating effect of local contemporaneous precipitation is significantly larger than for any other age group, and the coefficient on the interaction with upstream dam levels becomes significant when the control for precipitation is removed. If the effect of heat waves on mortality risk for the elderly is primarily driven by harvesting, i.e., displacing deaths in time that would counterfactually have occurred soon after instead of causing new deaths, upstream water availability on the heat-mortality relationship may only prevent deaths that would not otherwise be likely to occur. This is reasonable to expect given that water takes time to flow from upstream sources into a particular district for consumption, while precipitation is more immediate. However, without more specific data on the health status of elderly individuals upon their deaths and the causes of their deaths, it is not possible to estimate the ex-ante risk of death for these individuals, which is necessary to disentangle harvesting from causing deaths of otherwise healthy individuals. Thus further research is needed to determine why potable water availability does not significantly reduce the mortality risk of the elderly during heat waves.

B.1.1 Examining Heterogeneity by Household Water Source with DHS Data

The mortality records used to produce Tables 12-23 are simply monthly death counts by district, sex, age group, and population group, and do not include any additional information on each decedent included in those counts. Thus additional data is needed to estimate heterogeneous treatment effects by other household characteristics, such as the location of the household's primary water source. To accomplish this, I use Demographic and Health Survey (DHS) data from 2016 in South Africa to construct a measure of child mortality that can be linked to household-level covariates. The DHS includes records of up to 20 births for each surveyed woman with the birth date, current survival status, and, if applicable, the age upon death of each birthed child. Thus this data can be used to reconstruct child mortality rates over time for many years before the survey was administered. Prior literature has employed a similar strategy using DHS data to analyze mortality rates over time in sub-Saharan Africa (Garenne and Gakusi 2006).

The result of this reconstruction is a panel of 549 deaths of children under 5 years of age from 1998 to 2015, with the month of death, DHS sampling cluster location, and primary water source of the household for each child. There are 462 DHS sampling clusters with at least one child death in the sample, and despite the exact location being displaced to protect the privacy of DHS respondents, the location of each DHS cluster is substantially more precisely identified than the location of each decedent in Tables 12-23. Thus, in addition to enabling analysis of heterogeneity by household water source, using this data allows for the construction of much finer-grained measures of heat exposure and water availability, thereby replicating the results of Tables 12-23 with reduced spatial uncertainty. To this end, I reconstruct the measures of heat exposure and water availability according to the formulas in Section 2.3 using the displaced coordinates of each DHS sampling cluster.

Table 24 presents Poisson regression estimates using this panel of child death counts and the DHS-level measures of heat exposure and water availability. In column 1, the dependent variable includes children under 5 who were reported as siblings of a DHS respondent rather than children of a respondent, while column 2 excludes siblings. This is because the location of each sibling at their time of death is not provided in the data, so it is uncertain whether or not the sibling died in the same DHS sampling cluster as the respondent's primary residence. Columns 3 and 4 replicate columns 1 and 2 for children under 1 at their times of death to obtain a more exact replication of column 1 of Table 23.

In all columns, the mitigating effect of upstream water availability on mortality risk during heat waves is exclusive to households with a water source on-premises (about 62.4% of the sample of households reporting at least one child death). This suggests that, for infants and children, upstream water availability only reduces the risk of mortality during heat waves if that water does not need to be retrieved from outside the home for consumption. Most of these children are too young to be tasked with retrieving water by themselves, so this heterogeneity is more likely to arise indirectly via the health status of the mother. As before, I do not observe the health status of the child upon their death or the cause of their death, so further research with more specific data is needed to pin down mechanisms to explain this with certainty. In particular, an analysis of whether or not upstream water availability mitigates negative effects of heat exposure on prenatal health, and for whom, would shed light on why on-premises access to potable water is a prerequisite for this mitigating effect to reduce infant mortality risk.

B.2 Treatment Effects Conditional on Local Precipitation

The results in Tables 12-23 show that the estimated mitigating effect of upstream water availability on heat-related mortality is robust to the inclusion of precipitation controls. This means that the estimated effect of upstream water availability on the heat-mortality relationship cannot be explained by contemporaneous precipitation. However, this does not rule out the possibility of an interaction. For example, if upstream water availability and local precipitation are potable water source substitutes, the effect of upstream water availability on the heat-mortality relationship will be concentrated in periods of sparse local precipitation.

Figure 37 graphically represents the interaction between CDD 75°F and upstream dam levels with respect to the mortality rate, conditional on higher and lower precipitation respectively. When the precipitation in a district, as measured by the lagged 12-month mean of monthly total precipitation, is more abundant, there is no significant interaction between CDD 75°F and upstream dam levels. The key result of Section 2.4, the significant negative interaction between upstream dam levels and the slope of the heat-mortality relationship, only arises when precipitation is below the district's historical average. The size of the effect depicted in Figure 37(b) (a marginal effect of CDD 75°F of 0.05, 0.025, and 0 deaths per million at low, medium, and high levels of water availability, respectively) is much smaller than the size of the coefficient in Table 12 because, as depicted in Figure 5, the marginal effects of temperature on mortality increase above 70°F. Thus using CDD 75°F as a measure of heat exposure instead of CDD 90°F results in smaller coefficients when using a linear estimator. Figure 37 also demonstrates that the distributions of upstream dam levels conditional on higher and lower precipitation, respectively, are comparable, which confirms that the measure of water availability used in this paper is independent of local precipitation and its associated confounders.

The finding that upstream water availability and local precipitation are substitutes with respect to their effect on heat-related mortality risk is especially relevant to the climate change policy implications of this paper. Precipitation is projected to become more sparse and less predictable as climate change progresses (Allen et al. 2014), especially in the semi-arid climate of South Africa (Nkhonjera 2017), which already has spatially heterogeneous rainfall and a history of droughts. As this occurs, the availability of water from other sources will determine the degree to which heat-related mortality risk increases as a consequence. The measure of potable water availability used in this paper is one particular example of a source that is independent of local precipitation, but it is unlikely that the estimated effect is driven by anything specific to the process of water flowing downstream through naturally-occurring rivers and streams. I confirm this in Section B.3 using the Lesotho Highlands Water Project, a transnational water transfer, as a natural experiment increasing potable water availability from upstream sources in receiving districts through man-made means.

B.3 The Lesotho Highlands Water Project (LHWP)

In this section, I exploit a transnational water transfer inaugurated in 2004 as a natural experiment increasing upstream water availability in receiving districts. This serves two main functions. One, it is a robustness check to the results of section 4, which imply hot-season mortality should decline as a result of the water transfer. Two, it is an example of large-scale investment in water infrastructure. Thus, if the water transfer reduces heat-related mortality, it is evidence that investment in water infrastructure can be an effective community-level adaptation to heat.

The most densely populated province and industrial center of South Africa, Gauteng, is built on rocky, water-scarce, and high-elevation land. To meet the region's water needs, the government of South Africa and kingdom of Lesotho negotiated an agreement that led to the Lesotho Highlands Water Project (LHWP), the largest water transfer in the country. LHWP enables bilateral trade in which Lesotho diverts purchased water from the Senqunyane and Malibamat'so rivers toward the Vaal Dam in South Africa, and in return, South Africa generates hydroelectric power that is transferred back to Lesotho. The project was officially inaugurated in 2004, after which South Africa began purchasing water from Lesotho under the terms of the agreement. Once it reaches the Vaal Dam, a portion of the water is transferred to the cities of Johannesburg and Tshwane (Pretoria) in Gauteng for consumption. The remainder is stored in the Vaal Dam or released downstream, where it continues down the Vaal river to meet the Orange river at its confluence. For a detailed review of LHWP, including the reasons for its construction and political context, see Hitchcock (2012).

By diverting water from the highlands of Lesotho downstream to South Africa in a specific direction, LHWP constitutes a natural experiment that increased the supply of water in treated districts from upstream. As illustrated in the map in figure 38, I classify a district as "treated" by LHWP if it is either part of Gauteng Province or positioned downstream of the Vaal Dam along the Vaal and Orange rivers. In Figure 39, I confirm that twelve-month minimum dam levels increased in treated districts after the inauguration of LHWP. I then estimate the following difference-in-difference regression model described in equation 8. Let τ_j be an indicator variable that equals 1 if district j is treated by LHWP and zero otherwise.

$$
\begin{aligned}
M_{jt} = \beta_0^{LHWP} + \beta_\tau \tau_j + \beta_{CDD}^{LHWP} CDD_{jt} + \beta_{2004} \mathbf{1}_{y \geq 2004} + \gamma_{\tau,2004} \left(\tau_j \times \mathbf{1}_{y \geq 2004}\right) + \\
\gamma_{\tau,CDD} \left(\tau_j \times CDD_{jt}\right) + \gamma_{CDD,2004} \left(\mathbf{1}_{y \geq 2004} \times CDD_{jt}\right) + \phi_\tau \left(\tau_j \times CDD_{jt} \times \mathbf{1}_{y \geq 2004}\right) \\
+ \Omega \left(j \times y\right) + \epsilon_{jt}
\end{aligned} \tag{8}
$$

In equation 8, the coefficient of interest to the main hypothesis is ϕ_τ, which will be negative if the increase in upstream water supply resulting from LHWP reduced the slope of the heat-mortality relationship in treated districts. Figure 40 shows the hot-season mortality rate in treated and untreated districts over time with local regression smoothing (LOESS) curves to show trends. The figure provides evidence in favor of the necessary assumption of parallel trends before 2004, and the average death rates appear roughly equal across treated and non-treated districts. The parallel-trends test is presented explicitly in Table 25. After 2004, the average hot-season mortality rate in treated districts becomes significantly lower. While Figure 40 does not account for differences in excess heat incidence across years, the hypothesized difference-in-difference appears to be confirmed. The comparison in Figure 40 confirms that the magnitude of this difference does not emerge outside of the summer.

As a robustness check to further verify that the difference between treated and untreated groups is caused by LHWP's effect on upstream water availability, I also estimate a quadruple-difference model that interacts all terms in equation 8 with an indicator that equals 1 when the high-elevation water availability is above the 25th percentile and 0 otherwise. Since the dams supplying LHWP need to be high enough to divert water for LHWP to have a significant effect on upstream water availability, the coefficient on the quadruple-interaction term should be negative, and ϕ_τ itself should no longer be distinguishable from zero.

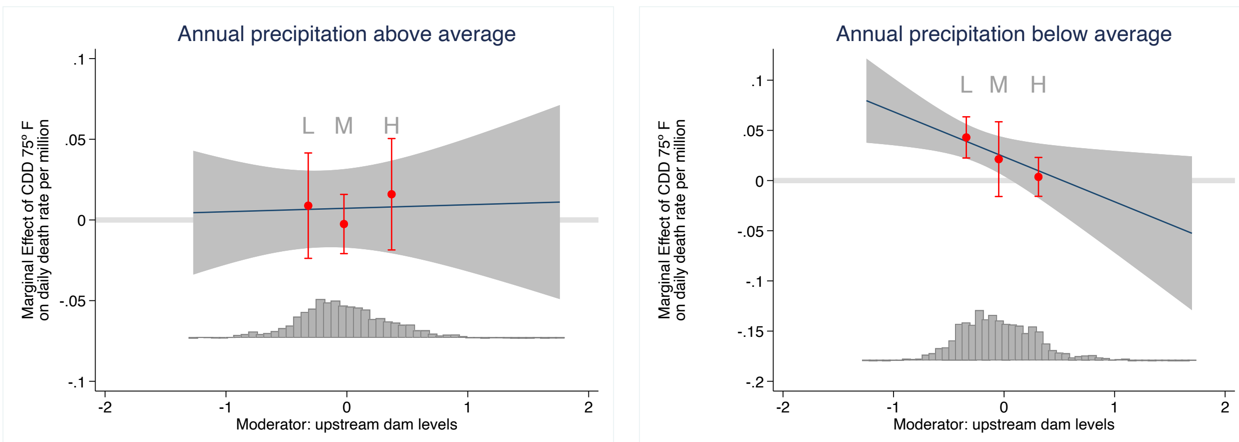
Table 26 provides the results of estimating equation 8 and the extended quadruple-difference model on the sample. The estimated ϕ_t is negative and statistically significant at the $\alpha = 0.1$ level. When the quadruple-interaction term with the high-elevation dam level indicator is introduced, ϕ_t loses statistical significance, and the quadruple-interaction term is negative and significant at the $\alpha = 0.05$ level. In sum, Table 26 confirms that the slope of the heat-mortality relationship differentially declined in treated districts following the inauguration of LHWP, and this differential decline was conditional on the source dam levels being high enough to supply the transfer with water. The size of the coefficients in Table 26 are similar to the coefficients in Table 28 (which shows the main specification described in Section 2.3.3 with CDD 75°F as the measure of heat exposure), suggesting that LHWP reduced the marginal effect of CDD 75°F in receiving districts by about 0.05 per million. This coefficient is much smaller than the coefficient in Table 12 because of the nonlinear effects of temperature on mortality. Since the marginal effect of temperature on mortality is increasing above 70°F, as depicted in Figure 5, using CDD 75°F instead of CDD 90°F results in smaller coefficients using a linear estimator.

B.4 Additional Figures

Figure 37: Effects of Upstream Water Availability on the Heat-Mortality Relationship Conditional on Contemporaneous Local Precipitation

(a) Higher precipitation

(b) Lower precipitation



Notes: Graphical representation of the interaction between upstream dam levels and CDD 75°F achieved with the Stata `interflex` package (Hainmueller et al. 2016). "L," "M," and "H" refer to the respective medians of pre-specified bins of upstream dam levels with cutoffs at the 33rd and 66th percentiles (low, medium, and high). The line and associated confidence interval represents the conventional linear multiplicative interaction estimator. The distribution of upstream dam levels conditional on higher and lower precipitation respectively is displayed below the interaction estimates in each graph. CDD 75°F is used in this figure instead of CDD 90°F to ensure an adequate number of nonzero observations in each bin to obtain meaningful estimates with the binning estimator.

Figure 38: Map of the Lesotho Highlands Water Project

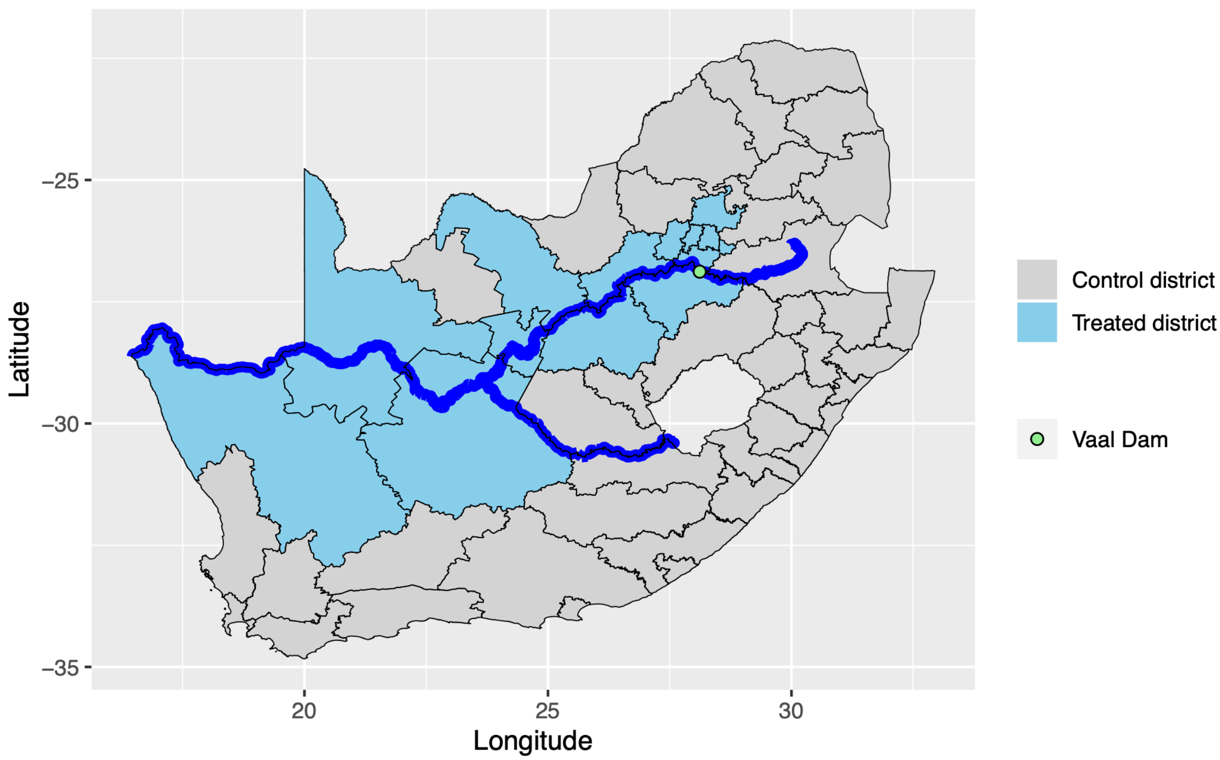


Figure 39: Minimum Dam Levels Before and After LHWP Treatment

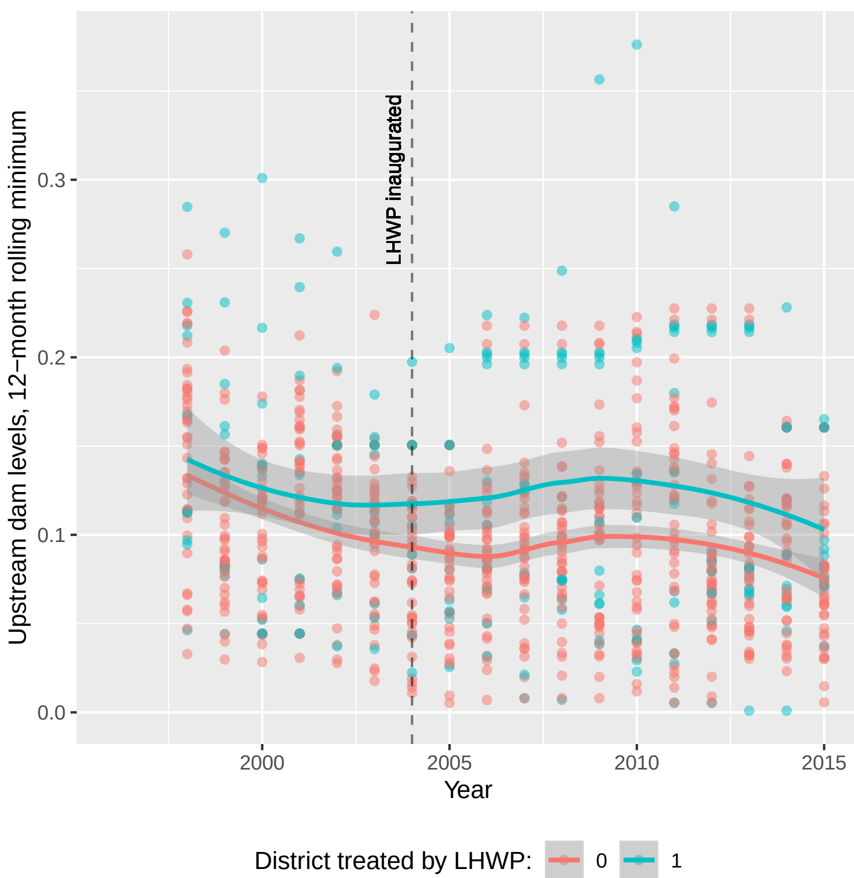


Figure 40: Trends in Hot-Season Mortality Before and After LHWP Treatment

(a) Summer

(b) Non-summer

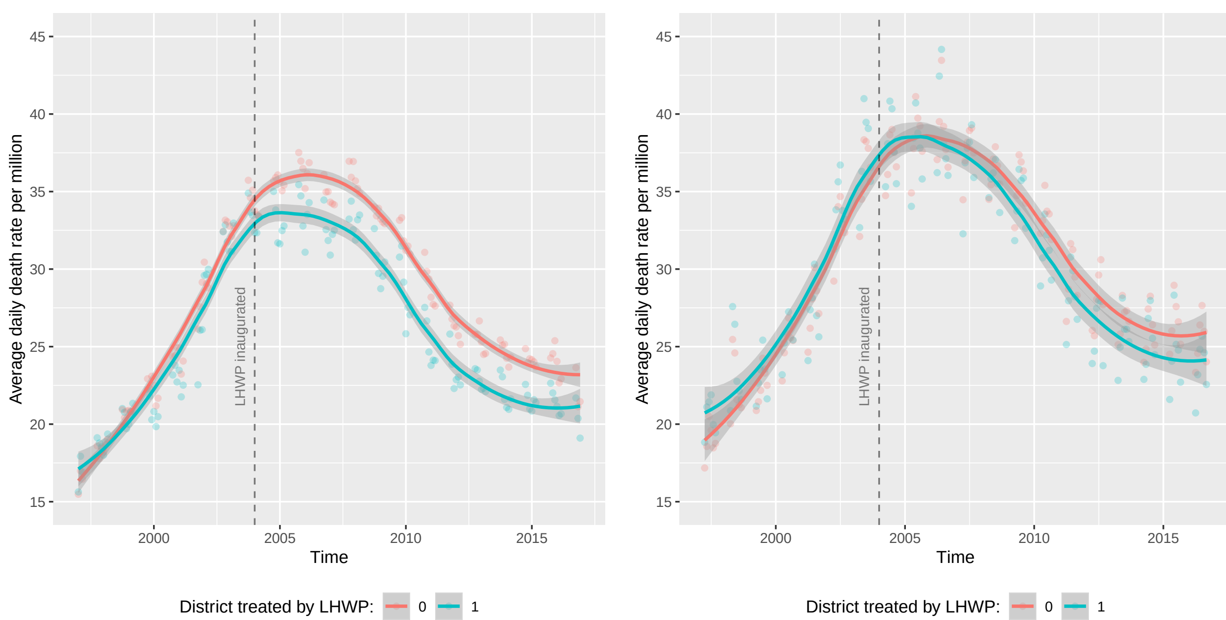
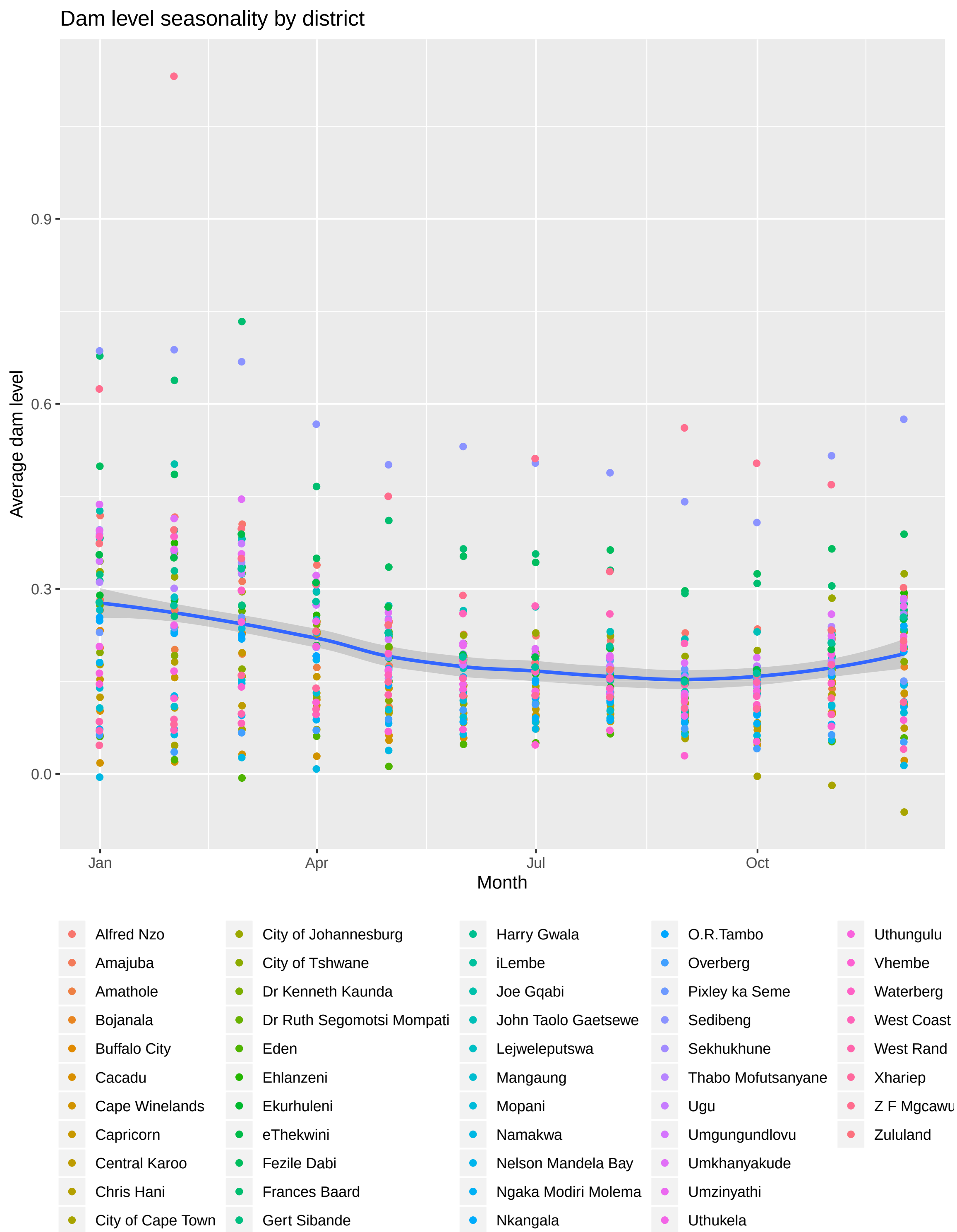


Figure 41: Seasonality of Dam Levels by District



B.5 Additional Tables

Table 21: Sex-Specific Effects of Water Availability on the Heat-Mortality Relationship

	Dependent variable: average daily deaths per million		
	(1)	(2)	(3)
	Men	Women	All
Upstream dam level × CDD 90°F	-9.23^{***}	-13.78^{***}	-9.23^{***}
	(1.36)	(1.17)	(1.36)
Female × Upstream dam level × CDD 90°F			-4.56^{***}
			(1.30)
Female × CDD 90°F			4.04^{***}
			(0.89)
Female × Upstream dam level			1.04
			(1.21)
Upstream dam level	-4.73^{***}	-3.70^{***}	-4.73^{***}
	(1.16)	(1.23)	(1.16)
CDD base temp 90°F	−0.26	3.84^{***}	−0.26
	(0.89)	(1.15)	(0.89)
Mean of dep. var.	65.07	53.09	59.20
N	88960	85685	174645
R^2	0.921	0.913	0.918

Notes: * p < 0.1, ** p < 0.05, *** p < 0.01. District-clustered standard errors in parentheses. "Average daily deaths per million" is the monthly death rate per million divided by the number of days in the month to address the fact that, all else equal, more deaths occur during months with more days. Dam levels are distance-weighted 12-month lagged averages calculated with equations 2 and 3. A dam is classified as upstream (downstream) of a district if it is at least 100 kilometers away from and 100 meters higher (lower) in elevation than the geographic center of the district. A dam is classified as within-district if it is within 100 kilometers and ± 100 meters elevation of the geographic center of the district. For graphical examples, see figures 7 and 8. Wherever within-district or downstream dam level controls are included, their interaction with CDD is also included. Cooling degree days (CDD) are a monthly incidence measure of heat above the base temperature. For reference, 1 CDD at a base temperature of 90°F is equivalent to 1 day in a month during which the average outside temperature exceeded 90°F by 1°F. The number of observations in this table is based on the number of population-subgroup-by-month observations, where each "population subgroup" is a possible permutation of sex, age group, and race. District-subgroup-year fixed effects, which absorb idiosyncratic annual level variation in mortality for each subgroup within each district, as well as controls for within-district dam levels, downstream dam levels, and monthly precipitation are included in all columns.

Table 22: Population Group-Specific Effects of Water Availability on the Heat-Mortality Relationship

	Dependent variable: average daily deaths per million			
	(1)	(2)	(3)	(4)
	Black African	Coloured	Asian/Indian	White
Upstream dam level × CDD 90°F	-12.37^{**}	-20.56^{**}	-0.48	-5.38^{***}
	(5.00)	(9.38)	(0.98)	(1.43)
Upstream dam level	-12.01^{***}	-0.45	-0.22	-2.01^{**}
	(2.94)	(1.18)	(0.24)	(0.76)
CDD base temp 90°F	2.49	3.09	1.05^{***}	-0.25
	(1.51)	(2.17)	(0.39)	(0.45)
Mean of dep. var.	153.22	29.15	4.76	23.91
District-subgroup-year FE	Yes	Yes	Yes	Yes
Within-district dam level control	Yes	Yes	Yes	Yes
Downstream dam level control	Yes	Yes	Yes	Yes
Monthly precipitation control	Yes	Yes	Yes	Yes
N	50676	46040	33006	44923
R^2	0.927	0.766	0.814	0.894

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. District-clustered standard errors in parentheses. "Average daily deaths per million" is the monthly death rate per million divided by the number of days in the month to address the fact that, all else equal, more deaths occur during months with more days. Dam levels are distance-weighted 12-month lagged averages calculated with equations 2 and 3. A dam is classified as upstream (downstream) of a district if it is at least 100 kilometers away from and 100 meters higher (lower) in elevation than the geographic center of the district. A dam is classified as within-district if it is within 100 kilometers and ± 100 meters elevation of the geographic center of the district. For graphical examples, see figures 7 and 8. Wherever within-district or downstream dam level controls are included, their interaction with CDD is also included. Cooling degree days (CDD) are a monthly incidence measure of heat above the base temperature. For reference, 1 CDD at a base temperature of 90°F is equivalent to 1 day in a month during which the average outside temperature exceeded 90°F by 1°F. The number of observations in this table is based on the number of population-subgroup-by-month observations, where each "population subgroup" is a possible permutation of sex, age group, and race. "District-subgroup-year fixed effects" absorb idiosyncratic annual level variation in mortality for each subgroup within each district.

Table 23: Age-Specific Effects of Water Availability on the Heat-Mortality Relationship

	Dep. var.: average daily deaths per million				
	(1)	(2)	(3)	(4)	(5)
	Infant	1−14	15−44	45−64	65+
Upstream dam level × CDD 90°F	−49.52***	−0.55***	−1.56***	−4.89**	−2.24
	(8.47)	(0.19)	(0.41)	(1.91)	(1.41)
Upstream dam level	−14.72***	−0.26***	−0.47**	−1.25**	−6.05***
	(4.50)	(0.10)	(0.22)	(0.47)	(1.64)
CDD base temp 90°F	9.34**	0.21*	0.33**	0.41	1.11*
	(4.59)	(0.11)	(0.15)	(0.42)	(0.57)
Mean of dep. var.	173.43	2.96	13.22	32.47	91.75
District-subgroup-year FE	Yes	Yes	Yes	Yes	Yes
Within-district dam control	Yes	Yes	Yes	Yes	Yes
Downstream dam control	Yes	Yes	Yes	Yes	Yes
Monthly precipitation control	Yes	Yes	Yes	Yes	Yes
N	27673	26688	36388	36429	36504
R^2	0.898	0.848	0.968	0.961	0.937

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. District-clustered standard errors in parentheses. "Average daily deaths per million" is the monthly death rate per million divided by the number of days in the month to address the fact that, all else equal, more deaths occur during months with more days. Dam levels are distance-weighted 12-month lagged averages calculated with equations 2 and 3. A dam is classified as upstream (downstream) of a district if it is at least 100 kilometers away from and 100 meters higher (lower) in elevation than the geographic center of the district. A dam is classified as within-district if it is within 100 kilometers and ± 100 meters elevation of the geographic center of the district. For graphical examples, see figures 7 and 8. Wherever within-district or downstream dam level controls are included, their interaction with CDD is also included. Cooling degree days (CDD) are a monthly incidence measure of heat above the base temperature. For reference, 1 CDD at a base temperature of 90°F is equivalent to 1 day in a month during which the average outside temperature exceeded 90°F by 1°F. "Downstream control" indicates whether or not downstream dam levels and their interaction with CDD were included as covariates in the regression model. The number of observations in this table is based on the number of population-subgroup-by-month observations, where each "population subgroup" is a possible permutation of sex, age group, and race. "District-subgroup-year fixed effects" absorb idiosyncratic annual level variation in mortality for each subgroup within each district.

Table 24: Poisson Regression Estimates with Reconstructed Child Mortality Rates from DHS Data

	Dependent variable: Count of deaths			
	(1)	(2)	(3)	(4)
	Under 5	Under 5	Under 1	Under 1
CDD base temp 75ºF × Upstream dam levels × Household water source on-premises	-0.15^{***}	-0.17^{***}	-0.14^{***}	-0.18^{***}
	(0.06)	(0.05)	(0.05)	(0.06)
CDD base temp 75ºF × Household water source on-premises	-0.008	-0.01	-0.005	-0.01
	(0.009)	(0.009)	(0.009)	(0.009)
CDD base temp 75ºF × Upstream dam levels	-0.02	-0.02	-0.04	-0.02
	(0.06)	(0.06)	(0.06)	(0.06)
Upstream dam levels × Household water source on-premises	3.59	3.31	4.53^{**}	4.04
	(2.18)	(2.47)	(2.26)	(2.58)
CDD base temp 75ºF	0.002	0.0006	0.002	0.0001
	(0.006)	(0.005)	(0.006)	(0.006)
Upstream dam levels	-1.44	-0.22	-2.32	-0.52
	(1.83)	(1.98)	(1.95)	(2.13)
Household water source on-premises	0.03	0.14	-0.05	0.09
	(0.20)	(0.20)	(0.21)	(0.21)
Siblings included	Yes	No	Yes	No
DHS sampling cluster fixed-effects	Yes	Yes	Yes	Yes
Year fixed-effects	Yes	Yes	Yes	Yes
Month fixed-effects	Yes	Yes	Yes	Yes
N (full)	34,558	30,366	29,614	26,710
N (non-zero dependent variable)	526	444	421	369
Pseudo R^2	0.072	0.039	0.075	0.043

Clustered (DHSID) standard-errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 25: Parallel Trends Tests for Lesotho Highlands Water Project (LHWP)

Difference-in-Difference

	Dependent variable: average daily deaths per million			
	(1)	(2)	(3)	(4)
	1997-2003	1997-2003	1997-2003	1997-2003
District treated by LHWP ×	−0.25		−0.32	
Time trend	(0.32)		(0.35)	
District treated by LHWP ×		−0.29		−1.14
1998		(2.38)		(2.59)
District treated by LHWP ×		−0.79		−1.48
1999		(2.38)		(2.60)
District treated by LHWP ×		−1.59		−1.31
2000		(2.38)		(2.55)
District treated by LHWP ×		−1.92		−2.22
2001		(2.38)		(2.58)
District treated by LHWP ×		−1.26		−2.32
2002		(2.38)		(2.60)
District treated by LHWP ×		−1.33		−1.96
2003		(2.38)		(2.64)
District treated by LHWP ×			−0.00	
Time trend × CDD 75° F			(0.02)	
District treated by LHWP ×				0.07
1998 × CDD 75° F				(0.16)
District treated by LHWP ×				−0.02
1999 × CDD 75° F				(0.16)
District treated by LHWP ×				0.01
2000 × CDD 75° F				(0.16)
District treated by LHWP ×				−0.03
2001 × CDD 75° F				(0.16)
District treated by LHWP ×				0.09
2002 × CDD 75° F				(0.16)
District treated by LHWP ×				−0.01
2003 × CDD 75° F				(0.15)
Observations	1,421	1,421	812	812
Joint signif. test P-value	0.53	0.98	0.97	0.98

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. "District treated by LHWP" means that a district received water from LHWP, i.e., was either in Gauteng province or positioned downstream of the Vaal Dam along the Vaal river basin.

Table 26: DiD and DDD Effects of Lesotho Highlands Water Project (LHWP) on Heat-Mortality Relationship in Treated Districts

	Dependent variable: daily death rate per million	
	(1)	(2)
Treated × After LHWP inauguration × CDD 75°F	-0.05^{***}	-0.01
	(0.02)	(0.03)
Dam levels near Lesotho above 25th percentile × Treated × After LHWP inauguration × CDD 75°F		-0.06^{**}
		(0.03)
Treated × CDD 75°F	0.08^{***}	0.07^{**}
	(0.02)	(0.03)
Treated × Dam levels near Lesotho above 25th percentile		-0.89
		(0.69)
Dam levels near Lesotho above 25th percentile × CDD 75°F		-0.04^{*}
		(0.02)
Dam levels near Lesotho above 25th percentile × Treated × CDD 75°F		0.01
		(0.03)
After LHWP inauguration × CDD 75°F	0.05^{***}	0.00
	(0.01)	(0.02)
CDD base temp 75°F	-3.41^{**}	14.80^{***}
	(1.53)	(3.72)
Mean of dep. var.	33.31	34.01
District-year FE	Yes	Yes
N	5833	5526
R^2	0.977	0.977

Notes: * p < 0.1, ** p < 0.05, *** p < 0.01. District-clustered standard errors in parentheses. "Average daily deaths per million" is the monthly death rate per million divided by the number of days in the month to address the fact that, all else equal, more deaths occur during months with more days. Dam levels are distance-weighted 12-month lagged averages calculated with equations 2 and 3. A dam is classified as "near Lesotho" if it is within 200 kilometers of either of the two dams serving the Lesotho Highlands Water Project included in the hydrological data from South Africa's Department of Water and Sanitation. Cooling degree days (CDD) are a monthly incidence measure of heat above the base temperature. For reference, 1 CDD at a base temperature of 75°F is equivalent to 1 day in a month during which the average outside temperature exceeded 75°F by 1°F.

Table 27: Data Description

Data	Source	Notes
Daily dam and reservoir levels (% of design capacity) at 773 monitoring stations throughout South Africa, 1996−2016	Department of Water and Sanitation (DWS) Hydrological Services	Aggregated to monthly means
Administrative death counts by district, month, population group, sex, and age group, 1997−2016	Statistics South Africa	
Monthly cooling degree days (CDD) at base temperatures of 90°F and 75°F at 68 weather stations throughout South Africa, 1997−2016	NOAA Global Historical Climatology Network (GHCN-Daily)	
Elevation extracted from NASA Shuttle Radar Topography Mission (SRTM) 3 arc-second (90m) raster data	Consortium for Spatial Information (CGIAR-CSI)	Covers entirety of South Africa and Lesotho
General Household Survey data, 2004 −2016	Statistics South Africa	Only available at the province level, thus not used in regressions

Table 28: Heat-Mortality Relationship Above 75°F Interacted with Dam Levels

	Dependent variable: average daily deaths per million			
	(1)	(2)	(3)	(4)
	Oct-Mar	Oct-Mar	Oct-Mar	All months
Upstream dam level × CDD 75°F	−0.02	−0.03***	−0.02**	
	(0.01)	(0.01)	(0.01)	
Upstream dam level × Summer indicator				−0.78**
				(0.30)
CDD base temp 75°F	0.01	0.01	0.01	
	(0.01)	(0.01)	(0.01)	
Upstream dam level	−0.85***	−0.04	−0.01	0.84
	(0.22)	(0.31)	(0.35)	(0.62)
Within-district dam level × CDD 75°F		0.02	0.02	
		(0.01)	(0.01)	
Within-district dam level × Summer indicator				0.11
				(0.23)
Within-district dam level		−1.13***	−1.11***	−0.80**
		(0.33)	(0.39)	(0.37)
Downstream dam level × CDD 75°F			−0.02	
			(0.02)	
Downstream dam level × Summer indicator				0.88***
				(0.27)
Downstream dam level			−0.05	−1.36**
			(0.44)	(0.54)
Mean of dep. var.	34.12	34.12	34.12	35.70
District-month-of-year FE	No	No	No	Yes
Within-district dam level control	No	Yes	Yes	Yes
Downstream dam level control	No	No	Yes	Yes
Monthly precipitation control	No	No	Yes	Yes
N	5474	5474	5474	11090
R^2	0.977	0.977	0.977	0.984

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. District-clustered standard errors in parentheses. "Average daily deaths per million" is the monthly death rate per million divided by the number of days in the month to address the fact that, all else equal, more deaths occur during months with more days. Dam levels are distance-weighted 12-month lagged averages calculated with equations 2 and 3. A dam is classified as upstream (downstream) of a district if it is at least 100 kilometers away from and 100 meters higher (lower) in elevation than the geographic center of the district. A dam is classified as within-district if it is within 100 kilometers and ± 100 meters elevation of the geographic center of the district. For graphical examples, see figures 7 and 8. Wherever within-district or downstream dam level controls are included, their interaction with CDD is also included. Cooling degree days (CDD) are a monthly incidence measure of heat above the base temperature. District-year FE are included in all columns.

Table 29: Heat-Mortality Relationship Above 75°F (Within-District Standard Deviations) Interacted with Dam Levels

	Dependent variable: average daily deaths per million		
	(1)	(2)	(3)
Upstream dam level × Std. CDD 75°F	-0.26^{***}	-0.37^{***}	-0.31^{**}
	(0.08)	(0.13)	(0.15)
Std. CDD base temp 75°F	-0.02	-0.02	0.09
	(0.07)	(0.07)	(0.09)
Upstream dam level	-0.86^{***}	-0.10	-0.02
	(0.21)	(0.31)	(0.35)
Within-district dam level × Std. CDD 75°F		0.15	0.17
		(0.12)	(0.14)
Within-district dam level		-1.03^{***}	-0.94^{***}
		(0.28)	(0.29)
Downstream dam level × Std. CDD 75°F			-0.09
			(0.18)
Downstream dam level			-0.25
			(0.36)
Mean of dep. var.	34.12	34.12	34.12
District-year FE	Yes	Yes	Yes
N	5474	5474	5474
R^2	0.977	0.977	0.977

Notes: * p < 0.1, ** p < 0.05, *** p < 0.01. District-clustered standard errors in parentheses. "Average daily deaths per million" is the monthly death rate per million divided by the number of days in the month to address the fact that, all else equal, more deaths occur during months with more days. Dam levels are distance-weighted 12-month lagged averages calculated with equations 2 and 3. A dam is classified as upstream (downstream) of a district if it is at least 100 kilometers away from and 100 meters higher (lower) in elevation than the geographic center of the district. A dam is classified as within-district if it is within 100 kilometers and ± 100 meters elevation of the geographic center of the district. For graphical examples, see figures 7 and 8. Wherever within-district or downstream dam level controls are included, their interaction with CDD is also included. Cooling degree days (CDD) are a monthly incidence measure of heat above the base temperature. For reference, 1 CDD at a base temperature of 75°F is equivalent to 1 day in a month during which the average outside temperature exceeded 75°F by 1°F.

Table 30: Heat-Mortality Relationship Above 90°F Interacted with Dam Levels (Excluding Within-District)

	Dependent variable: log(avg daily deaths per million)			
	(1)	(2)	(3)	(4)
	Oct-Mar	Oct-Mar	Oct-Mar	All months
Upstream dam level × CDD 90°F	−0.87***	−0.90***	−0.83***	
	(0.17)	(0.18)	(0.28)	
Upstream dam level × Summer indicator				−0.78**
				(0.30)
CDD base temp 90°F	0.80***	0.81***	0.91***	
	(0.25)	(0.26)	(0.15)	
Upstream dam level	−1.02***	−0.49	−0.49	0.63
	(0.20)	(0.30)	(0.30)	(0.60)
Downstream dam level × CDD 90°F		0.00	−0.13	
		(0.29)	(0.34)	
Downstream dam level × Summer indicator				0.95***
				(0.26)
Downstream dam level		−0.89**	−0.89**	−1.89***
		(0.34)	(0.34)	(0.57)
Mean of dep. var.	34.12	34.12	34.12	35.37
District-year FE	Yes	Yes	Yes	Yes
District-month-of-year FE	No	No	No	Yes
Monthly precipitation control	No	No	Yes	Yes
N	5474	5474	5474	11649
R^2	0.977	0.977	0.977	0.984

Notes: * p < 0.1, ** p < 0.05, *** p < 0.01. District-clustered standard errors in parentheses. "Average daily deaths per million" is the monthly death rate per million divided by the number of days in the month to address the fact that, all else equal, more deaths occur during months with more days. Dam levels are distance-weighted 12-month lagged averages calculated with equations 2 and 3. A dam is classified as upstream (downstream) of a district if it is at least 100 kilometers away from and 100 meters higher (lower) in elevation than the geographic center of the district. A dam is classified as within-district if it is within 100 kilometers and ± 100 meters elevation of the geographic center of the district. For graphical examples, see figures 7 and 8. Wherever within-district or downstream dam level controls are included, their interaction with CDD is also included. Cooling degree days (CDD) are a monthly incidence measure of heat above the base temperature. For reference, 1 CDD at a base temperature of 90°F is equivalent to 1 day in a month during which the average outside temperature exceeded 90°F by 1°F.

Table 31: Heat-Mortality Relationship Above 90°F Interacted with Dam Levels (Including Month Fixed Effects)

	Dependent variable: average daily deaths per million		
	(1)	(2)	(3)
Upstream dam level × CDD 90°F	-0.65^{**}	-0.68^{**}	-0.68^{*}
	(0.28)	(0.28)	(0.38)
CDD base temp 90°F	0.77^{***}	0.77^{**}	0.78^{***}
	(0.28)	(0.35)	(0.26)
Upstream dam level	-0.05	0.15	0.15
	(0.30)	(0.32)	(0.33)
Within-district dam level × CDD 90°F		0.02	0.02
		(0.64)	(0.68)
Within-district dam level		-0.48^{**}	-0.47^{*}
		(0.22)	(0.26)
Downstream dam level × CDD 90°F			-0.02
			(0.44)
Downstream dam level			-0.04
			(0.37)
Mean of dep. var.	34.12	34.12	34.12
District-year FE	Yes	Yes	Yes
Month FE	Yes	Yes	Yes
Monthly precipitation control	No	No	Yes
N	5474	5474	5474
R^2	0.981	0.981	0.981

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. District-clustered standard errors in parentheses. "Average daily deaths per million" is the monthly death rate per million divided by the number of days in the month to address the fact that, all else equal, more deaths occur during months with more days. Dam levels are distance-weighted 12-month lagged averages calculated with equations 2 and 3. A dam is classified as upstream (downstream) of a district if it is at least 100 kilometers away from and 100 meters higher (lower) in elevation than the geographic center of the district. A dam is classified as within-district if it is within 100 kilometers and ± 100 meters elevation of the geographic center of the district. For graphical examples, see figures 7 and 8. Wherever within-district or downstream dam level controls are included, their interaction with CDD is also included. Cooling degree days (CDD) are a monthly incidence measure of heat above the base temperature. For reference, 1 CDD at a base temperature of 90°F is equivalent to 1 day in a month during which the average outside temperature exceeded 90°F by 1°F.

Appendix C - The Regressive Costs of Drinking Water Contaminant Avoidance

C.1 Alternative Data Source: Nielsen Retail Scanner Data

In this section, I report the results of the panel fixed-effects regression approach described in Section 3.2.2 using the Nielsen Retail Scanner dataset in place of the HomeScan Consumer Panel. The main advantage of using this dataset is that, unlike the Consumer Panel, it records transactions directly from retailers and thus should provide a more complete and precise representation of consumption in a particular area. However, because no information is provided on the purchaser, I cannot distinguish between purchases made by lower- and higher-income households respectively. As a result, in place of the "Below 200% FPL" indicator variable used in Section 3.3, I use a continuous variable indicating the percentage of individuals in a county who are below 200% of the federal poverty line in a given year according to the 5-year American Community Survey, which was available from 2009 to present. The results presented in Table use the Retail Scanner data from 2009 to 2016, which spans 2,671 counties.

The results in Table 32 are consistent with the results in Table 13 from the main text in the estimated effects on total Other expenditure (negative and marginally not significant, $p \approx 0.104$) and nutrition values (negative and significant across the board). Table 33 discretizes the poverty measure into quartiles, demonstrating that the negative effects on nutrition values are strongest (in terms of both magnitude and statistical significance) among observations in the top quartile of poverty rates. However, the interaction coefficient for water expenditure is not statistically significant, nor does the point estimate have the expected sign, providing more evidence in favor of considering alternative mechanisms for the effect on calories such as health-related income effects and avoidance of foods that require water to prepare in-home.

These results share many of the limitations described in Section 3.4. Like the panel data, the scanner data does not contain information on households' restaurant takeout and fast food purchases, meaning it is not possible to distinguish between overall declines in food consumption and substitution of food away from home for grocery purchases. Additionally, because treatment is assigned at the county level and the location of stores reporting data to Nielsen is only specified at the first-three-digits-of-ZIP-code level (similar to metro area), it is not possible to ascertain whether the stores reflected in the scanner data are in the neighborhood(s) affected by each violation. However, as before, these limitations are likely to work against finding a significant effect.

C.2 Additional Tables

Table 32: Differential Effects of Active Health-Based Water Quality Violations (Scanner Data)

	Water exp. (2004 USD)	Other exp. (2004 USD)	Calories (Cal)	Fat (Grams)	Carbs (Grams)	Sugars (Grams)	Protein (Grams)
Active Violation × Pct. Below 200% FPL	-0.001	-0.076	-1.335**	-0.056**	-0.171**	-0.088**	-0.035*
	(0.001)	(0.047)	(0.618)	(0.027)	(0.076)	(0.035)	(0.018)
Active Violation	0.007	0.806	14.070**	0.579**	1.835**	0.946**	0.351*
	(0.011)	(0.523)	(6.689)	(0.291)	(0.814)	(0.373)	(0.199)
Pct. Below 200% FPL	0.001	-0.112	-2.746*	-0.086	-0.368**	-0.162**	-0.069*
	(0.002)	(0.0100)	(1.462)	(0.061)	(0.182)	(0.080)	(0.040)
Mean of dep. var.	0.43	36.76	410.62	16.75	51.38	23.79	10.94
N	149868	149868	149868	149868	149868	149868	149868
Adjusted R^2	0.883	0.946	0.942	0.936	0.944	0.942	0.930

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. County-clustered standard errors are in parentheses. All variables in this table are in per-capita terms. Expenditure measures were deflated to 2004 USD using the annual food-at-home CPI measures from USDA (Kuhns et al. 2015). County and panel month fixed effects are included in all columns. Regressions were weighted using the method proposed by Imai and Kim (2019) to mitigate bias in two-way fixed effects (TWFE) linear regressions.

Table 33: Differential Effects of Active Health-Based Water Quality Violations (Scanner Data)

	Water exp. (2004 USD)	Other exp. (2004 USD)	Calories (Cal)	Fat (Grams)	Carbs (Grams)	Sugars (Grams)	Protein (Grams)
Active Violation × Second Poverty Quartile	-0.004	-0.238	-4.868	-0.143	-0.771	-0.392	-0.071
	(0.008)	(0.397)	(5.068)	(0.217)	(0.626)	(0.289)	(0.149)
Active Violation × Third Poverty Quartile	0.000	-0.492	-4.708	-0.153	-0.756	-0.447	-0.077
	(0.009)	(0.415)	(5.501)	(0.240)	(0.671)	(0.310)	(0.160)
Active Violation × Fourth Poverty Quartile	-0.006	-0.641	-11.419**	-0.465**	-1.530**	-0.805***	-0.268*
	(0.011)	(0.411)	(5.414)	(0.236)	(0.658)	(0.302)	(0.158)
Active Violation	0.002	0.356	5.535	0.194	0.831*	0.444*	0.097
	(0.008)	(0.330)	(4.091)	(0.182)	(0.489)	(0.227)	(0.125)
Second Poverty Quartile	0.011	0.017	1.646	0.113	0.119	0.036	0.078
	(0.007)	(0.357)	(4.489)	(0.185)	(0.570)	(0.259)	(0.127)
Third Poverty Quartile	0.008	-0.447	-8.536	-0.245	-1.235	-0.499	-0.192
	(0.010)	(0.436)	(5.879)	(0.238)	(0.759)	(0.332)	(0.167)
Fourth Poverty Quartile	0.018	-0.302	-8.658	-0.189	-1.302	-0.493	-0.178
	(0.013)	(0.608)	(8.612)	(0.360)	(1.080)	(0.474)	(0.238)
Mean of dep. var.	0.43	36.76	410.62	16.75	51.38	23.79	10.94
N	149868	149868	149868	149868	149868	149868	149868
Adjusted R^2	0.882	0.946	0.941	0.935	0.944	0.941	0.930

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. County-clustered standard errors are in parentheses. All variables in this table are in per-capita terms. Expenditure measures were deflated to 2004 USD using the annual food-at-home CPI measures from USDA (Kuhns et al. 2015). County and panel month fixed effects are included in all columns. Regressions were weighted using the method proposed by Imai and Kim (2019) to mitigate bias in two-way fixed effects (TWFE) linear regressions.

Table 34: Differential Effects of Active Health-Based Water Quality Violations Based on Duration (Full Sample)

	Water exp. (2004 USD)	Other exp. (2004 USD)	Calories (Cal)	Fat (Grams)	Carbs (Grams)	Sugars (Grams)	Protein (Grams)
Active Violation (Month 7+)	0.132**	-2.328	-42.154***	-1.670***	-5.201***	-3.073***	-1.101**
× Below 200% FPL	(0.061)	(1.767)	(13.779)	(0.643)	(1.752)	(0.961)	(0.442)
Active Violation (Month 7+)	0.008	0.116	-4.156	-0.211	-0.305	0.138	-0.214**
	(0.017)	(0.631)	(3.277)	(0.147)	(0.424)	(0.243)	(0.102)
Active Violation (Month 1-6)	0.139***	0.529	-13.601	-0.681	-1.391	-1.570*	-0.649*
× Below 200% FPL	(0.051)	(2.084)	(10.208)	(0.491)	(1.402)	(0.896)	(0.332)
Active Violation (Month 1-6)	-0.005	-1.147**	-5.225**	-0.153	-0.607**	-0.427***	-0.112
	(0.013)	(0.457)	(2.298)	(0.114)	(0.282)	(0.164)	(0.069)
Below 200% FPL	-0.055	-3.820***	-88.995***	-3.756***	-11.234***	-4.986***	-1.971***
	(0.047)	(1.346)	(11.313)	(0.498)	(1.383)	(0.676)	(0.331)
Mean of dep. var.	2.03	204.30	1233.78	50.96	152.66	68.19	32.14
P-value: Short vs. long	0.882	0.153	0.055	0.150	0.049	0.171	0.353
N	4328963	4328963	4328963	4328963	4328963	4328963	4328963
Adjusted R^2	0.374	0.564	0.430	0.417	0.390	0.313	0.438

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. County-clustered standard errors are in parentheses. "Full Sample" refers to the complete sample included in Table 13, which includes all households who experienced at least one water quality violation (regardless of duration) during the study period. Expenditure measures were deflated to 2004 USD using the annual food-at-home CPI measures from USDA (Kuhns et al. 2015). Household and panel month fixed effects are included in all columns. Regressions were weighted using the method proposed by Imai and Kim (2019) to mitigate bias in two-way fixed effect (TWFE) linear regressions. The row "P-value: short vs. long" reports statistical comparison tests between the coefficient for "Active Violation (Month 7+) × Below 200% FPL" and the coefficient for "Active Violation (Month 1-6) × Below 200% FPL."

Bibliography

Achyuta Adhvaryu, James Fenske, and Anant Nyshadham. Early Life Circumstance and Adult Mental Health. *Journal of Political Economy*, 127(4):1516–1549, August 2019. ISSN 00223808.

Kenneth R. Ahern, Ran Duchin, and Tyler Shumway. Peer Effects in Risk Aversion and Trust. *Review of Financial Studies*, 27(11):3213–3240, November 2014. ISSN 08939454.

Maura Allaire, Haowei Wu, and Upmanu Lall. National trends in drinking water quality violations. *Proceedings of the National Academy of Sciences*, 115(9):2078–2083, February 2018. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1719805115. Publisher: National Academy of Sciences Section: Social Sciences.

Maura Allaire, Taylor Mackay, Shuyan Zheng, and Upmanu Lall. Detecting community response to water quality violations using bottled water sales. *Proceedings of the National Academy of Sciences*, 116(42):20917–20922, October 2019. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1905385116. Publisher: National Academy of Sciences Section: Social Sciences.

Hunt Allcott and Todd Rogers. The Short-Run and Long-Run Effects of Behavioral Interventions: Experimental Evidence from Energy Conservation. *American Economic Review*, 104(10):3003–3037, October 2014. ISSN 0002-8282. doi: 10.1257/aer.104.10.3003.

Myles R. Allen, Vicente R. Barros, John Broome, Wolfgang Cramer, Renate Christ, John A. Church, Leon Clarke, Qin Dahe, Purnamita Dasgupta, and Navroz K. Dubash. IPCC fifth assessment synthesis report-climate change 2014 synthesis report. 2014.

F.H. Allport. *Social Psychology.* Riverside Press, Cambridge, MA, 1924.

Patricia M. Anderson, Kristin F. Butcher, Hilary W. Hoynes, and Diane Whitmore Schanzenbach. Beyond Income: What Else Predicts Very Low Food Security Among Children? *Southern Economic Journal*, 82(4):1078–1105, 2016. ISSN 2325-8012. doi: 10.1002/soej.12079.

Chon K. Ao. *Essays on the impact of clean water on human capital and productivity.* PhD Thesis, 2016.

Cemal Eren Arbatlı, Quamrul H. Ashraf, Oded Galor, and Marc Klemp. Diversity and Conflict. *Econometrica*, 88(2):727–797, 2020. ISSN 1468-0262. doi: https://doi.org/10.3982/ECTA13734. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA13734.

Dan Ariely, Uri Gneezy, George Loewenstein, and Nina Mazar. Large Stakes and Big Mistakes. *Review of Economic Studies*, 76(2):451–469, April 2009. ISSN 00346527, 1467937X. doi: 10.1111/j.1467-937X.2009.00534.x.

Philip Babcock, Kelly Bedard, Gary Charness, John Hartman, and Heather Royer. Letting down the Team? Social Effects of Team Incentives. *Journal of the European Economic Association*, 13(5):841–870, 2015. ISSN 1542-4774. doi: https://doi.org/10.1111/jeea.12131. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jeea.12131.

Rakesh Banerjee and Riddhi Bhowmick. Heat, Infant Mortality and Adaptation: Evidence from India. 2016.

Alan Barreca, Karen Clay, Olivier Deschênes, Michael Greenstone, and Joseph S. Shapiro. Convergence in Adaptation to Climate Change: Evidence from High Temperatures and Mortality, 1900–2004. *American Economic Review*, 105(5):247–251, 2015.

Alan Barreca, Karen Clay, Olivier Deschenes, Michael Greenstone, and Joseph S. Shapiro. Adapting to Climate Change: The Remarkable Decline in the US Temperature-Mortality Relationship over the Twentieth Century. *Journal of Political Economy*, 124(1):105–159, 2016.

Xavier Basagaña, Claudio Sartini, Jose Barrera-Gómez, Payam Dadvand, Jordi Cunillera, Bart Ostro, Jordi Sunyer, and Mercedes Medina-Ramón. Heat Waves and Cause-specific Mortality at all Ages. *Epidemiology*, 22(6):765–772, 2011. ISSN 10443983.

Irina Basieva, Emmanuel Pothos, Jennifer Trueblood, Andrei Khrennikov, and Jerome Busemeyer. Quantum probability updating from zero priors (by-passing Cromwell's rule). *Journal of Mathematical Psychology*, 77:58–69, April 2017. ISSN 0022-2496. doi: 10.1016/j.jmp.2016.08.005.

William F. Bassett and Robin L. Lumsdaine. Probability Limits: Are Subjective Assessments Adequately Accurate? *Journal of Human Resources*, 36(2):327–363, 2001. ISSN 0022166X.

Brian Beach, Joseph Ferrie, Martin Saavedra, and Werner Troesken. Typhoid Fever, Water Quality, and Human Capital Formation. *The Journal of Economic History*, 76(01):41–75, March 2016.

Avner Ben-Ner, Brian P. McCall, Massoud Stephane, and Hua Wang. Identity and in-group/out-group differentiation in work and giving behaviors: Experimental evidence. *Journal of Economic Behavior & Organization*, 72(1):153–170, October 2009. ISSN 0167-2681. doi: 10.1016/j.jebo.2009.05.007.

Laurent Bergé. Efficient estimation of maximum likelihood models with multiple fixed-effects: the R package FENmlm. *CREA Discussion Papers*, (13), 2018.

Spiros Bougheas, Jeroen Nieboer, and Martin Sefton. Risk-taking in social settings: Group and peer effects. *Journal of Economic Behavior & Organization*, 92:273–283, August 2013. ISSN 0167-2681. doi: 10.1016/j.jebo.2013.06.010.

Roy Brouwer, Fumbi Crescent Job, Bianca van der Kroon, and Richard Johnston. Comparing Willingness to Pay for Improved Drinking-Water Quality Using Stated Preference Methods in Rural and Urban Kenya. *Applied Health Economics and Health Policy*, 13(1):81–94, February 2015. ISSN 11755652.

James A. Brox, Ramesh C. Kumar, and Kenneth R. Stollery. Estimating Willingness to Pay for Improved Water Quality in the Presence of Item Nonresponse Bias. *American Journal of Agricultural Economics*, 85(2):414–428, May 2003. ISSN 00029092.

Robin Burgess, Olivier Deschenes, Dave Donaldson, and Michael Greenstone. Weather, climate change and death in India. *University of Chicago*, 2017.

Marshall Burke and Kyle Emerick. Adaptation to Climate Change: Evidence from US Agriculture. *American Economic Journal: Economic Policy*, 8(3):106–140, 2016. ISSN 1945-7731.

Maria Paula Cacault and Manuel Grieder. How group identification distorts beliefs. *Journal of Economic Behavior & Organization*, 164:63–76, August 2019. ISSN 0167-2681. doi: 10.1016/j.jebo.2019.05.027.

Donald T. Campbell. Common fate, similarity, and other indices of the status of aggregates of persons as social entities. *Behavioral Science*, 3 (1):14–25, 1958. ISSN 1099-1743. doi: 10.1002/bs.3830030103. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bs.3830030103.

Alexander W. Cappelen, Ulrik H. Nielsen, Bertil Tungodden, Jean-Robert Tyran, and Erik Wengstrom. Fairness Is Intuitive. *Experimental Economics*, 19(4):727–740, December 2016. ISSN 13864157.

Alexander W. Cappelen, Trond Halvorsen, Erik O. Sorensen, and Bertil Tungodden. Face-Saving or Fair-Minded: What Motivates Moral Behavior? *Journal of the European Economic Association*, 15(3):540–557, July 2017. ISSN 15424766.

Andrea Carlson, Elina T. Page, Thea Palmer Zimmerman, Carina E. Tornow, and Sigurd Hermansen. Linking USDA Nutrition Databases to IRI Household-Based and Store-Based Scanner Data, 2019.

Fortuna Casoria, Ernesto Reuben, and Christina Rott. The Effect of Group Identity on Hiring Decisions With Incomplete Information. SSRN Scholarly Paper ID 3731536, Social Science Research Network, Rochester, NY, November 2020.

Simone Cerroni. Eliciting farmers' subjective probabilities, risk, and uncertainty preferences using contextualized field experiments. *Agricultural Economics*, 51 (5):707–724, 2020. ISSN 1574-0862. doi: 10.1111/agec.12587. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/agec.12587.

Ian Chadd, Emel Filiz-Ozbay, and Erkut Y. Ozbay. The relevance of irrelevant information. *Experimental Economics*, 24(3):985–1018, September 2021. ISSN 1573-6938. doi: 10.1007/s10683-020-09687-3.

Lopamudra Chakraborti. Do plants' emissions respond to ambient environmental quality?

Evidence from the clean water act. *Journal of Environmental Economics and Management*, 79:55–69, September 2016. ISSN 0095-0696. doi: 10.1016/j.jeem.2016.04.005.

Gary Charness and Matthew Rabin. Understanding Social Preferences with Simple Tests. *The Quarterly Journal of Economics*, 117(3):817–869, August 2002. ISSN 0033-5533. doi: 10.1162/003355302760193904.

Nicole Chavez. Michigan will end Flint's free bottled water program, 2018. Library Catalog: www.cnn.com.

Daniel L. Chen, Martin Schonger, and Chris Wickens. oTree—An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9:88–97, March 2016. ISSN 2214-6350. doi: 10.1016/j.jbef.2015.12.001.

Yan Chen and Sherry Xin Li. Group Identity and Social Preferences. *American Economic Review*, 99(1):431–457, March 2009. ISSN 0002-8282. doi: 10.1257/aer.99.1.431.

Raj Chetty, John N Friedman, Nathaniel Hendren, Maggie R Jones, and Sonya R Porter. The Opportunity Atlas: Mapping the Childhood Roots of Social Mobility. Working Paper 25147, National Bureau of Economic Research, October 2018. Series: Working Paper Series.

Abbigail J. Chiodo. Subjective Probabilities: Psychological Theories and Economic Applications. *Federal Reserve Bank of St. Louis Review*, 86(1):33–47, January 2004. ISSN 00149187.

Peter Christensen, David Keiser, and Gabriel Lade. Economic Effects of Environmental Crises: Evidence from Flint, Michigan. *SSRN Electronic Journal*, 2019. ISSN 1556-5068. doi: 10.2139/ssrn.3420526.

Karen Clay, Werner Troesken, and Michael Haines. Lead and Mortality. *The Review of Economics and Statistics*, 96(3):458–470, 2014. ISSN 0034-6535. Publisher: The MIT Press.

Frank C. Curriero, Karlyn S. Heiner, Jonathan M. Samet, Scott L. Zeger, Lisa Strug, and Jonathan A. Patz. Temperature and Mortality in 11 Cities of the Eastern United States. *American Journal of Epidemiology*, 155(1):80–87, January 2002. ISSN 0002-9262. doi: 10.1093/aje/155.1.80.

David Danz, Lise Vesterlund, and Alistair J. Wilson. Belief Elicitation: Limiting Truth Telling with Information on Incentives. Working Paper 27327, National Bureau of Economic Research, June 2020. Series: Working Paper Series.

Lucas W. Davis and Paul J. Gertler. Contribution of air conditioning adoption to future energy use under global warming. *Proceedings of the National Academy of Sciences*, 112 (19):5962–5967, May 2015. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1423558112.

Jonathan de Quidt, Johannes Haushofer, and Christopher Roth. Measuring and Bounding

Experimenter Demand. *American Economic Review*, 108(11):3266–3302, November 2018. ISSN 0002-8282. doi: 10.1257/aer.20171330.

Adeline Delavande and Basit Zafar. Stereotypes and Madrassas: Experimental evidence from Pakistan. *Journal of Economic Behavior & Organization*, 118:247–267, October 2015. ISSN 0167-2681. doi: 10.1016/j.jebo.2015.03.020.

Melissa Dell, Benjamin F. Jones, and Benjamin A. Olken. What Do We Learn from the Weather? The New Climate-Economy Literature. *Journal of Economic Literature*, 52(3): 740–798, September 2014. ISSN 0022-0515. doi: 10.1257/jel.52.3.740.

Stefano DellaVigna and Devin Pope. What Motivates Effort? Evidence and Expert Forecasts. *The Review of Economic Studies*, 85(2):1029–1069, April 2018. ISSN 0034-6527. doi: 10.1093/restud/rdx033.

Olivier Deschenes. Temperature, human health, and adaptation: A review of the empirical literature. *Energy Economics*, 46:606–619, November 2014. ISSN 0140-9883. doi: 10.1016/j.eneco.2013.10.013.

Olivier Deschênes and Michael Greenstone. The Economic Impacts of Climate Change: Evidence from Agricultural Output and Random Fluctuations in Weather. *The American Economic Review*, 97(1):354–385, 2007.

Olivier Deschênes, Michael Greenstone, and Jonathan Guryan. Climate Change and Birth Weight. *American Economic Review*, 99(2):211–217, May 2009. ISSN 00028282.

Florencia Devoto, Esther Duflo, Pascaline Dupas, William Parienté, and Vincent Pons. Happiness on Tap: Piped Water Adoption in Urban Morocco. *American Economic Journal: Economic Policy*, 4(4):68–99, 2012.

Salvatore Di Falco, Marcella Veronesi, and Mahmud Yesuf. Does Adaptation to Climate Change Provide Food Security? A Micro-Perspective From Ethiopia. *American Journal of Agricultural Economics*, 93(3):829–846, 2011. ISSN 0002-9092.

Thomas Dohmen, David Huffman, Jürgen Schupp, Armin Falk, Uwe Sunde, and Gert G. Wagner. Individual Risk Attitudes: Measurement, Determinants, and Behavioral Consequences. *Journal of the European Economic Association*, 9(3):522–550, 2011. ISSN 1542-4766. Publisher: Oxford University Press.

Esther Duflo and Rohini Pande. Dams. *The Quarterly Journal of Economics*, 122(2):601–646, 2007. ISSN 0033-5533.

Liran Einav, Ephraim Leibtag, and Aviv Nevo. Recording discrepancies in Nielsen Homescan data: Are they present and do they matter? *QME*, 8(2):207–239, June 2010. ISSN 1573-711X. doi: 10.1007/s11129-009-9073-0.

Andrew Ellis and Michele Piccione. Correlation Misperception in Choice. *American Economic Review*, 107(4):1264–1292, April 2017. ISSN 0002-8282. doi: 10.1257/aer.20160093.

Willis D. Ellis, editor. *A source book of Gestalt psychology.* A source book of Gestalt psychology. Kegan Paul, Trench, Trubner & Company, London, England, 1938. doi: 10.1037/11496-000. Pages: xiv, 403.

Benjamin Enke. What You See Is All There Is. *The Quarterly Journal of Economics*, 135 (3):1363–1398, August 2020. ISSN 0033-5533. doi: 10.1093/qje/qjaa012.

Benjamin Enke and Florian Zimmermann. Correlation Neglect in Belief Formation. *The Review of Economic Studies*, 86(1):313–332, January 2019. ISSN 0034-6527. doi: 10.1093/restud/rdx081.

Larry G. Epstein and Jiankang Zhang. Subjective Probabilities on Subjectively Unambiguous Events. *Econometrica*, 69(2):265–306, March 2001. ISSN 00129682.

Ernst Fehr and Klaus M. Schmidt. A Theory of Fairness, Competition, and Cooperation. *The Quarterly Journal of Economics*, 114(3):817–868, August 1999. ISSN 0033-5533. doi: 10.1162/003355399556151.

Joseph P. Ferrie and Werner Troesken. Water and Chicago's mortality transition, 1850-1925. *Explorations in Economic History*, 45(1):1–16, 2008.

Melissa L. Finucane, Paul Slovic, C. K. Mertz, James Flynn, and Theresa A. Satterfield. Gender, race, and perceived risk: The 'white male' effect. *Health, Risk & Society*, 2(2): 159–172, July 2000. ISSN 1369-8575. doi: 10.1080/713670162. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/713670162.

Raphael Flepp. Uninformative Performance Signals and Forced CEO Turnover. SSRN Scholarly Paper ID 3904056, Social Science Research Network, Rochester, NY, August 2021.

James Flynn, Paul Slovic, and C. K. Mertz. Gender, Race, and Perception of Environmental Health Risks. *Risk Analysis*, 14(6):1101–1108, 1994. ISSN 1539-6924. doi: https://doi.org/10.1111/j.1539-6924.1994.tb00082.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1539-6924.1994.tb00082.x.

Richard Freeman, Wenquan Liang, Ran Song, and Christopher Timmins. Willingness to pay for clean air in China. *Journal of Environmental Economics and Management*, 94: 188–216, March 2019. ISSN 0095-0696. doi: 10.1016/j.jeem.2019.01.005.

David E. Frisvold. Nutrition and Cognitive Achievement: An Evaluation of the School Breakfast Program. *Journal of Public Economics*, 124:91–104, April 2015. ISSN 00472727.

Alberto Gallace and Charles Spence. To what extent do Gestalt grouping principles influence tactile perception? *Psychological Bulletin*, 137(4):538–561, July 2011. ISSN 0033-2909. doi: http://dx.doi.org/10.1037/a0022335. Num Pages: 538-561 Publisher: American Psychological Association (US).

Michel Garenne and Enéas Gakusi. Health transitions in sub-Saharan Africa: overview of mortality trends in children under 5 years old (1950-2000). *Bulletin of the World Health Organization*, 84 6:470–8, 2006.

Teevrat Garg, Stuart E. Hamilton, Jacob P. Hochard, Evan Plous Kresch, and John Talbot. (Not so) gently down the stream: River pollution and health in Indonesia. *Journal of Environmental Economics and Management*, 92:35–53, November 2018. ISSN 0095-0696. doi: 10.1016/j.jeem.2018.08.011.

Francesca Gioia. Incentive Schemes and Peer Effects on Risk Behaviour: An Experiment. *Theory and Decision*, 87(4):473–495, November 2019. ISSN 00405833.

Lorenz Goette, David Huffman, and Stephan Meier. The impact of group membership on cooperation and norm enforcement: Evidence using random assignment to real social groups. *American Economic Review*, 96(2):212–216, 2006.

Joshua Graff Zivin and Matthew Neidell. Temperature and the Allocation of Time: Implications for Climate Change. *Journal of Labor Economics*, 32(1):1–26, January 2014. ISSN 0734-306X. doi: 10.1086/671766.

Jay P. Graham, Mitsuaki Hirai, and Seung-Sup Kim. An Analysis of Water Collection Labor among Women and Children in 24 Sub-Saharan African Countries. *PLOS ONE*, 11(6): e0155981, June 2016. doi: 10.1371/journal.pone.0155981.

Veronika Grimm, Verena Utikal, and Lorenzo Valmasoni. In-Group Favoritism and Discrimination among Multiple Out-Groups. *Journal of Economic Behavior and Organization*, 143:254–271, November 2017. ISSN 01672681.

Francesco Guala and Antonio Filippin. The Effect of Group Identity On Distributive Choice: Social Preference or Heuristic? *The Economic Journal*, 127(602):1047–1068, June 2017. ISSN 0013-0133. doi: 10.1111/ecoj.12311.

Werner Güth, Matteo Ploner, and Tobias Regner. Determinants of in-group bias: Is group affiliation mediated by guilt-aversion? *Journal of Economic Psychology*, 30(5):814–827, October 2009. ISSN 0167-4870. doi: 10.1016/j.joep.2009.07.001.

Karel Haal, Anja Smith, and Eddy van Doorslaer. The rise and fall of mortality inequality in South Africa in the HIV era. *SSM - Population Health*, 5:239–248, June 2018. ISSN 2352-8273. doi: 10.1016/j.ssmph.2018.06.007.

Jens Hainmueller, Jonathan Mummolo, and Yiqing Xu. *How Much Should We Trust Estimates from Multiplicative Interaction Models? Simple Tools to Improve Empirical Practice.* 2016. URL `https://papers.ssrn.com/abstract_id=2739221`.

Shakoor Hajat and Tom Kosatky. Heat-related mortality: a review and exploration of heterogeneity. *Journal of epidemiology and community health*, 64(9):753, 2010.

Shakoor Hajat, Ben G. Armstrong, Nelson Gouveia, and Paul Wilkinson. Mortality Displacement of Heat-Related Deaths: A Comparison of Delhi, São Paulo, and London. *Epi-*

demiology, 16(5):613–620, 2005. ISSN 1044-3983.

Ruth Hamill, Timothy D. Wilson, and Richard E. Nisbett. Insensitivity to sample bias: Generalizing from atypical cases. *Journal of Personality and Social Psychology*, 39(4):578–589, 1980. ISSN 1939-1315(Electronic),0022-3514(Print). doi: 10.1037/0022-3514.39.4.578. Place: US Publisher: American Psychological Association.

Robert K. Hitchcock. The Lesotho Highlands Water Project: Water, Culture, and Environmental Change. In Barbara Rose Johnston, Lisa Hiwasaki, Irene J. Klaver, Ameyali Ramos Castillo, and Veronica Strang, editors, *Water, Cultural Diversity, and Global Environmental Change: Emerging Trends, Sustainable Futures?*, pages 319–338. Springer Netherlands, Dordrecht, 2012. ISBN 978-94-007-1774-9. doi: 10.1007/978-94-007-1774-9_23.

Tanjim Hossain and Ryo Okui. The Binarized Scoring Rule. *The Review of Economic Studies*, 80(3 (284)):984–1001, 2013. ISSN 0034-6527. Publisher: [Oxford University Press, The Review of Economic Studies, Ltd.].

Hilary Hoynes, Marianne Page, and Ann Huff Stevens. Can targeted transfers improve birth outcomes?: Evidence from the introduction of the WIC program. *Journal of Public Economics*, 95(7):813–827, August 2011. ISSN 0047-2727. doi: 10.1016/j.jpubeco.2010.12.006.

Hilary Hoynes, Doug Miller, and David Simon. Income, the Earned Income Tax Credit, and Infant Health. *American Economic Journal: Economic Policy*, 7(1):172–211, 2015. ISSN 1945-7731.

Hilary W. Hoynes and Diane Whitmore Schanzenbach. Consumption Responses to In-Kind Transfers: Evidence from the Introduction of the Food Stamp Program. *American Economic Journal: Applied Economics*, 1(4):109–139, 2009. ISSN 1945-7782.

Michael D. Hurd. Subjective Probabilities in Household Surveys. *Annual Review of Economics*, 1(1):543–562, 2009. ISSN 19411383.

Michael D. Hurd and Kathleen McGarry. The Predictive Validity of Subjective Probabilities of Survival. *Economic Journal*, 112(482):966–985, October 2002. ISSN 00130133.

Kosuke Imai and In Song Kim. When Should We Use Unit Fixed Effects Regression Models for Causal Inference with Longitudinal Data? *American Journal of Political Science*, 63(2):467–490, 2019. ISSN 1540-5907. doi: 10.1111/ajps.12417. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajps.12417.

Rhiannon L. Jerch. The Local Consequences of Federal Mandates: Evidence from the Clean Water Act. 2018.

Robert J. Johnston and Paul J. Thomassin. Willingness to Pay for Water Quality Improvements in the United States and Canada: Considering Possibilities for International Meta-analysis and Benefit Transfer. *Agricultural and Resource Economics Review*, 39(1):

114–131, February 2010. ISSN 10682805.

Matthew E. Kahn. The Climate Change Adaptation Literature. *Review of Environmental Economics and Policy*, 10(1):166–178, January 2016. ISSN 1750-6816. doi: 10.1093/reep/rev023.

Daniel Kahneman and Amos Tversky. Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3(3):430–454, July 1972. ISSN 0010-0285. doi: 10.1016/0010-0285(72)90016-3.

Ido Kallir and Doron Sonsino. The Neglect of Correlation in Allocation Decisions. *Southern Economic Journal*, 75(4):1045–1066, April 2009. ISSN 00384038.

Charles F. F. Karney. Algorithms for geodesics. *Journal of Geodesy*, 87(1):43–55, January 2013. ISSN 1432-1394. doi: 10.1007/s00190-012-0578-z.

Lance C. Keene, Jessica M. Dehlin, Jim Pickett, Kathryn R. Berringer, Iman Little, Ashley Tsang, Alida M. Bouris, and John A. Schneider. #PrEP4Love: success and stigma following release of the first sex-positive PrEP public health campaign. *Culture, Health & Sexuality*, 23(3):397–413, March 2021. ISSN 1369-1058. doi: 10.1080/13691058.2020.1715482. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/13691058.2020.1715482.

Lawrence Keketso. The Mixed Blessings of the Lesotho Highlands Water Project. *Mountain Research and Development*, 23(1):7–10, February 2003. ISSN 0276-4741, 1994-7151. doi: 10.1659/0276-4741(2003)023[0007:TMBOTL]2.0.CO;2.

Pascal Kieren and Martin Weber. Expectation Formation under Uninformative Signals. page 49, 2021.

Iliana V. Kohler, Collin F. Payne, Chiwoza Bandawe, and Hans-Peter Kohler. The Demography of Mental Health among Mature Adults in a Low-Income, High-HIV-Prevalence Context. *Demography*, 54(4):1529–1558, August 2017. ISSN 00703370.

Joachim I. Krueger and Theresa E. DiDonato. Social Categorization and the Perception of Groups and Group Differences. *Social and Personality Psychology Compass*, 2 (2):733–750, 2008. ISSN 1751-9004. doi: 10.1111/j.1751-9004.2008.00083.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1751-9004.2008.00083.x.

Annemarie Kuhns, Richard Volpe, Ephraim Leibtag, and Ed Roeger. How USDA Forecasts Retail Food Price Inflation, 2015.

Amrei M. Lahno and Marta Serra-Garcia. Peer Effects in Risk Taking: Envy or Conformity? *Journal of Risk and Uncertainty*, 50(1):73–95, February 2015. ISSN 08955646.

Gilat Levy and Ronny Razin. Correlation Neglect, Voting Behavior, and Information Aggregation. *American Economic Review*, 105(4):1634–1645, April 2015a. ISSN 00028282.

Gilat Levy and Ronny Razin. Does Polarisation of Opinions Lead to Polarisation of Platforms? The Case of Correlation Neglect. *Quarterly Journal of Political Science*, 10(3):

321–355, 2015b. ISSN 15540626.

Michelle Marcus. Testing the Water: Drinking Water Quality, Public Notification, and Child Outcomes. *Review of Economics and Statistics*, pages 1–45, 2020. Publisher: MIT Press.

Andrew J. McDermott and Mark B. Stephens. Cost of eating: whole foods versus convenience foods in a low-income model. *Family Medicine*, 42(4):280–284, April 2010. ISSN 1938-3800.

Stephan Meier, Lamar Pierce, Antonino Vaccaro, and Barbara La Cara. Trust and in-group favoritism in a culture of crime. *Journal of Economic Behavior & Organization*, 132:78–92, December 2016. ISSN 0167-2681. doi: 10.1016/j.jebo.2016.09.005.

Brian Mullen, John F Dovidio, Craig Johnson, and Carolyn Copper. In-group-out-group differences in social projection. *Journal of Experimental Social Psychology*, 28(5):422–440, September 1992. ISSN 0022-1031. doi: 10.1016/0022-1031(92)90040-Q.

Jonathan Mummolo and Erik Peterson. Demand Effects in Survey Experiments: An Empirical Assessment. *American Political Science Review*, 113(2):517–529, May 2019. ISSN 0003-0554, 1537-5943. doi: 10.1017/S0003055418000837. Publisher: Cambridge University Press.

Clayton Neighbors, Mary E. Larimer, and Melissa A. Lewis. Targeting misperceptions of descriptive drinking norms: efficacy of a computer-delivered personalized normative feedback intervention. *Journal of Consulting and Clinical Psychology*, 72(3):434–447, June 2004. ISSN 0022-006X. doi: 10.1037/0022-006X.72.3.434.

Clayton Neighbors, Amanda J. Dillard, Melissa A. Lewis, Rochelle L. Bergstrom, and Teryl A. Neil. Normative Misperceptions and Temporal Precedence of Perceived Norms and Drinking. *Journal of studies on alcohol*, 67(2):290–299, March 2006. ISSN 0096-882X.

Nicolas G. Nelson, Christy L. Collins, R. Dawn Comstock, and Lara B. McKenzie. Exertional Heat-Related Injuries Treated in Emergency Departments in the U.S., 1997–2006. *American Journal of Preventive Medicine*, 40(1):54–60, January 2011. ISSN 0749-3797. doi: 10.1016/j.amepre.2010.09.031.

Kristoffer P. Nimark and Savitar Sundaresan. Inattention and belief polarization. *Journal of Economic Theory*, 180:203–228, March 2019. ISSN 0022-0531. doi: 10.1016/j.jet.2018.12.007.

German K. Nkhonjera. Understanding the impact of climate change on the dwindling water resources of South Africa, focusing mainly on Olifants River basin: A review. *Environmental Science & Policy*, 71:19–29, 2017.

Axel Ockenfels and Peter Werner. Beliefs and Ingroup Favoritism. *Journal of Economic Behavior and Organization*, 108:453–462, December 2014. ISSN 01672681.

Fabian Paetzel and Rupert Sausgruber. Cognitive ability and in-group bias: An experimental study. *Journal of Public Economics*, 167:280–292, November 2018. ISSN 0047-2727. doi:

10.1016/j.jpubeco.2018.04.006.

Christina Palmer. Risk perception: Another look at the 'white male' effect. *Health, Risk & Society*, 5(1):71–83, March 2003. ISSN 1369-8575. doi: 10.1080/1369857031000066014. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/1369857031000066014.

Linda S. Perloff and Philip Brickman. False consensus and false uniqueness: Biases in perceptions of similarity. *Academic Psychology Bulletin*, 4(3):475–494, 1982. ISSN 0193-1709. Place: US Publisher: Michigan Psychological Assn.

Ilya Rahkovsky, Young Jo, and Andrea Carlson. Consumers Balance Time and Money in Purchasing Convenience Foods, 2018.

Patricia I. Ritter. The Hidden Role of Piped Water in the Prevention of Obesity in Developing Countries. Experimental and Non-Experimental Evidence. Technical Report 2019-02, University of Connecticut, Department of Economics, August 2019.

Lilia Rodriguez-Tapia, Daniel A. Revollo-Fernandez, and Jorge A. Morales-Novelo. Household's Perception of Water Quality and Willingness to Pay for Clean Water in Mexico City. *Economies*, 5(2):1–14, June 2017. ISSN 22277099.

Lee Ross, David Greene, and Pamela House. The false consensus effect: An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology*, 13(3):279–301, 1977. ISSN 1096-0465. doi: 10.1016/0022-1031(77)90049-X. Place: Netherlands Publisher: Elsevier Science.

Jonathan Roth, Pedro H C Sant'Anna, Alyssa Bilinski, and John Poe. What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature. page 54, 2022.

Michelle Saksena, Abigail Okrent, Tobenna D. Anekwe, Clare Cho, Chris Dicken, Anne Effl, Howard Elitzak, Joanne Guthrie, Karen Hamrick, Jeffrey Hyman, Young Jo, Biing-Hwan Lin, Lisa Mancino, Patrick W. McLaughlin, Ilya Rahkovsky, Katherine Ralston, Travis A. Smith, Hayden Stewart, Jessica E. Todd, and Charlotte Tuttle. America's Eating Habits: Food Away From Home, 2018. Library Catalog: www.ers.usda.gov.

Anna Sandberg. Competing Identities: A Field Study of In-group Bias Among Professional Evaluators. *The Economic Journal*, 128(613):2131–2159, 2018. ISSN 1468-0297. doi: https://doi.org/10.1111/ecoj.12513. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ecoj.12513.

Laurel A. Schaider, Lucien Swetschinski, Christopher Campbell, and Ruthann A. Rudel. Environmental justice and drinking water quality: are there socioeconomic disparities in nitrate levels in U.S. drinking water? *Environmental Health*, 18(1):3, January 2019. ISSN 1476-069X. doi: 10.1186/s12940-018-0442-6.

Wolfram Schlenker and Michael J. Roberts. Nonlinear temperature effects indicate severe damages to U.S. crop yields under climate change. *Proceedings of the National Academy* *of Sciences*, 106(37):15594–15598, September 2009. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0906865106.

C.R. Snyder and J. Shenkel. The "Illusion" of Uniqueness. 1978. doi: 10.1177/002216787801800305.

Alessandro Sontuoso, Cristina Bicchieri, Alexander Funcke, and Einav Hart. Strategic Problems with Risky Prospects: The Impact of Feedback in Complex Interactions. page 45, 2021.

Jerry Suls and C.K. Wan. In search of the false-uniqueness phenomenon: Fear and estimates of social consensus". *Journal of Personality and Social Psychology*, 52(1):211–217, 1987. doi: 10.1037/0022-3514.52.1.211. ISBN: 1939-1315.

David Switzer and Manuel P. Teodoro. Class, Race, Ethnicity, and Justice in Safe Drinking Water Compliance. *Social Science Quarterly*, 99(2):524–535, 2018. ISSN 1540-6237. doi: 10.1111/ssqu.12397. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ssqu.12397.

Henri Tajfel, M. G. Billig, R. P. Bundy, and Claude Flament. Social categorization and intergroup behaviour. *European Journal of Social Psychology*, 1 (2):149–178, 1971. ISSN 1099-0992. doi: 10.1002/ejsp.2420010202. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ejsp.2420010202.

Werner Troesken. *Water, Race, and Disease*. MIT Press, 2004. ISBN 978-0-262-20148-3. Google-Books-ID: GmgYTwfcIMsC.

Werner Troesken, Joseph P. Ferrie, and Karen Rolf. Cognitive Disparities, Lead Plumbing, and Water Chemistry: Intelligence Test Scores and Exposure to Water-Borne Lead Among World War Two U.S. Army Enlistees. Technical report, National Bureau of Economic Research, 2011.

John C. Turner, Michael A. Hogg, Penelope J. Oakes, Stephen D. Reicher, and Margaret S. Wetherell. *Rediscovering the social group: A self-categorization theory*. Rediscovering the social group: A self-categorization theory. Basil Blackwell, Cambridge, MA, US, 1987. ISBN 978-0-631-14806-7. Pages: x, 239.

Amos Tversky and Daniel Kahneman. Judgment under Uncertainty: Heuristics and Biases. *Science*, 185(4157):1124–1131, September 1974. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.185.4157.1124. Publisher: American Association for the Advancement of Science Section: Articles.

United States Environmental Protection Agency. Data Reliability Analysis of the EPA Safe Drinking Water Information System/Federal Version (SDWIS/FED), 2000. Library Catalog: nepis.epa.gov.

United States Environmental Protection Agency. Safe Drinking Water Act (SDWA) Resources and FAQs, February 2020. Library Catalog: echo.epa.gov.

Francesca Valent, D'Anna Little, Roberto Bertollini, Leda E Nemer, Fabio Barbone, and Giorgio Tamburlini. Burden of disease attributable to selected environmental factors and injury among children and adolescents in Europe. *The Lancet*, 363(9426):2032–2039, June 2004. ISSN 0140-6736. doi: 10.1016/S0140-6736(04)16452-0.

Johan Wagemans, James H. Elder, Michael Kubovy, Stephen E. Palmer, Mary A. Peterson, Manish Singh, and Rüdiger von der Heydt. A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure–ground organization. *Psychological Bulletin*, 138 (6):1172–1217, November 2012. ISSN 0033-2909. doi: http://dx.doi.org/10.1037/a0029333. Num Pages: 1172-1217 Publisher: American Psychological Association (US).

Max Wertheimer. Laws of organization in perceptual forms. In *A source book of Gestalt psychology*, pages 71–88. Kegan Paul, Trench, Trubner & Company, London, England, 1938. doi: 10.1037/11496-005.

Catherine Wolfram, Orie Shelef, and Paul Gertler. How Will Energy Demand Develop in the Developing World? *Journal of Economic Perspectives*, 26(1):119–138, February 2012. ISSN 0895-3309. doi: 10.1257/jep.26.1.119.

Joshua Graff Zivin, Matthew Neidell, and Wolfram Schlenker. Water Quality Violations and Avoidance Behavior: Evidence from Bottled Water Consumption. *American Economic Review*, 101(3):448–453, May 2011. ISSN 0002-8282. doi: 10.1257/aer.101.3.448.

Joshua S. Graff Zivin, Yingquan Song, Qu Tang, and Peng Zhang. Temperature and High-Stakes Cognitive Performance: Evidence from the National College Entrance Examination in China. Working Paper 24821, National Bureau of Economic Research, July 2018.