STATEMENT ABOUT PAST AND FUTURE RESEARCH: KEITH CHEN

I am a behavioral economist who studies individual choice behavior. While my work has ranged over a number of topics and subjects, I am drawn to questions where I believe that the prevailing explanation for an important human behavior is unnecessarily complex. I try to find simpler, more general drivers, located at a more basic level of explanation. I believe several of my projects shed new light on the way people make important economic decisions. I will briefly summarize the major areas of my current research, and locate my work in the broader literature. I've organized this statement to emphasize current topics I am working on, with explanations of older research — that I am no longer actively working on — after that.

Current Work: Using Smartphone Data to study Human Behavior and Cognition

Much of my most recent work attempts to utilize a new form of data — smartphone location information — to study what we can learn about people from their movement patterns. 81% of adult Americans own a smartphone, and the data they produce represents a tremendous opportunity to learn about cognition and decision making. Recently, my co-authors and I have conducted a series of studies that use anonymized data from more than 20 million smartphone users to study what we think is one of the most pressing issues facing cognitive science, understanding increasing political polarization and partisan cognition.

To understand these projects, it helps to understand the nature of smartphone location data. They are created by applications which ask permission to know the phone's location, even when the app in not in use (this is common for weather, navigation, and social apps). Several firms purchase and aggregate these data across applications to form a more complete picture of peoples' movement and interactions. These data consists of billions of "pings" per week, where each ping reports a smartphone's unique ID, a timestamp, a latitude and longitude, and an estimate of the accuracy of that location estimate. Smartphones "ping" at uncertain intervals, but the modal time between pings is 10 minutes, dropping to 5 seconds when the phone thinks it needs greater location accuracy (say, when driving). A movement data panel allows us to know where and when people spend time, and to a lesser extent, with whom.

The bulk of projects I have been working on with these data merge them with two data sets that my co-authors and I have constructed for the 2016 US presidential election: a set of latitude and longitude polygons (also called geofences) representing building rooftop outlines of US polling places, geofences for their associated voting precinct boundaries, and the official US vote counts for each of those precincts. By merging smartphone owner's movement data with the precincts in which they live, my coauthors and I have begun several projects which aim to understand how political identity and affiliations affect peoples' lives, and the ways partisanship affects both information processing and decision making.

The earliest of these projects examines Thanksgiving travel patterns as a lens on politically-driven family discord. We show that shortly following the historically divisive 2016 US presidential election, Thanksgiving dinners attended by opposing-party precinct residents were 30-50 minutes shorter than same-party gatherings. To account for potential confounds like rural-urban or regional differences, we show that our measured Thanksgiving declines survive (if anything, grow) when fixed-effects restrict comparisons to Thanksgiving travel whose host live within a mile of each other, as do both traveling families.

Two other comparisons suggest that this decline in Thanksgiving dinner durations is causally a result of partisan animosity. First, for a small set of diners we can further compare durations directly to the length of that same dinner in 2015, and find declines in length only for cross-partisan dinners. Second, dinner reductions in 2016 tripled for travelers from media markets with heavy political advertising — an effect also not observed in 2015 before ads were run — implying a relationship to partisanship which is amplified through election-related advertising. These estimates scale up to a total of 62 million Thanksgiving hours lost to post-election partisanship in 2016, many of which could have been prevented by reductions in political advertising. This first paper was recently published in *Science*, and was also recognized with an "editor's choice" award from the journal *Nature*.

This first paper set out to demonstrate that partisan cleavages appear to be distorting what many Americans think of as an important and private institution — family ties. In a closely related paper, we study whether partisan beliefs about "fake news" and government institutions influence not just public decisions like voting, but high-stakes personal decisions. To do so we study hurricane evacuations over the 2016 and 2017 North American hurricane seasons, which span both before and after the emergence and viral spread of "hurricane trutherism", conservative-media dismissals of hurricane warnings as politically motivated.

We conduct a difference-in-difference study of the evacuation behavior of conservative vs liberal coastal residents and find evidence that evacuating an impending hurricane has become a partisan issue. Clinton and Trump voters evacuated Hurricanes Matthew (October 2016) and Harvey (August 2017) at identical rates. In September 2017 (shortly after Harvey but before Irma struck Florida), prominent conservative commentators (most notably Rush Limbaugh and Ann Coulter) questioned the motivation behind hurricane advisories; suggesting that they were being overblow in order to highlight climate-change, and that residents shouldn't evacuate or overreact to Irma, which was being reported as the most energetic Pacific hurricane ever measured. Following this viral "hurricane trutherism", we find that 30% of Trump-voting Florida residents evacuated Irma compared to 44% of Clinton, a wedge which is unexplained by both demographic and geographic covariates. We see this work as contributing to our empirical understanding of partisan identity and cognitive effects; people appear to make even large life-or-death decisions through a partisan lens. This paper has been revised and resubmitted to PNAS.

My co-authors and I are continuing to work with these data to extract further insights from smartphone data for political outcomes. In our most recent working paper, we study smartphone-measured waiting times at 93,658 US polling place and find large racial disparities in how long Americans have to wait to vote. Relative to entirely-white neighborhoods, residents of entirely-black neighborhoods waited 29% longer to vote and were 74% more likely to spend more than 30 minutes waiting to vote. Two-thirds of this disparity persists even when comparing predominantly white and black polling places within the same county. Smaller but strongly significant effects hold for other demographics such as education and income. We hope that beyond simply measuring undesirable disparities in access to voting, this work provides policymakers a tool to address these disparities, which can be measured and monitored in an inexpensive, comprehensive, and scalable way. This paper is under review at Science.

Current Work: Labor Economics Using Data from Uber

Recently, I took a leave of absence from UCLA to work as the head of economics and research at Uber, a startup company who manage a large "gig" economy ridesharing marketplace. While there I mainly focused on algorithms for efficient marketplace intermediation, such as writing the "surge" pricing algorithm, and on managing the integration of different platforms which use a shared supply base: integrating pooled rides and food delivery with the main ride-sharing platform. While doing this though, I also worked on projects which use Uber data on rider and driver behavior as a lens on human decision making. Since leaving Uber I've begun to publish these projects.

The first of these projects focusses on how to both conceptualize and measure how much value people put on control of their own work schedule, i.e. work flexibility. Most traditional jobs require a worker to commit to a pre-determined work schedule, say 9-5 Monday through Friday. If we think of this as zero flexibility, many jobs display negative flexibility: many workers can be asked last minute to fill a shift they did not expect, and can be terminated for a repeated inability to do so. By contrast, many gig-economy jobs exhibit the opposite: positive work flexibility. Uber drivers, for example, can choose their schedules in real time: starting work whenever they want to and stopping whenever not on a trip. My coauthors and I develop a revealed-preference estimator that can both measure the value workers derive from this, and predict workers' responses to counterfactually less flexible work arrangements. We fit this model to both naturally occurring and experimentally induced wage variation for Uber drivers, and find that our model both suggests large amounts of worker surplus form work flexibility, and predicts out-of-sample labor decisions with great precision.

In a nutshell: what we estimate for Uber drivers is a reservation-wage process, fit so as to rationalize observed dynamic labor-supply decisions. Studying the behavior of drivers on Uber's main, Uber-X platform, we observe both an Uber driver's decision in any given hour to work or

not, and the prevailing wage in their city that they would be expected to earn if they chose to work. What we observe then is a series of inequality on their realized reservation wages, i.e. whether prevailing wages were high enough to induce them to work any given hour. Using a Bayesian mixture model, we estimate for each driver a mean reservation wage for every hour of the week (168), plus three variances of a weekly, daily, and hourly shock to this reservation wage. Simply put, in this model a driver works if and only if in a particular hour in their city, realized wages are higher than their mean reservation wage for this hour of the week plus three shocks: a shock specific to this hour, a daily shock that is common to all hours this day, and a weekly shock common to all hours this week. The size of these shocks is identified by the within-driver predictability of their work patterns at different levels of temporal resolution. Intuitively, hourly shocks are large (have a high variance) if, even if a worker works the same number of hours this Monday than they did last Monday, *which* hours they work moves around a lot in ways not explained by variation in hourly earnings. The values of flexibility comes from the ability of a worker to work only those hours where they will earn in excess of their reservation wage, rather than having to choose hours in larger blocks or shifts for which this is only true on average. Intuitively then, the larger these shocks the more flexibility matters to a driver, and the relative size of these shocks describes at which temporal level flexibility matters most.

Our results indicate that while the Uber relationship may have other drawbacks, Uber drivers benefit significantly from real-time flexibility, earning more than twice the surplus they would in less flexible arrangements. Counterfactually, if drivers were required to supply labor inflexibly at prevailing wages, they would reduce the hours they supply by more than two-thirds. In additional work, we explore which types of people appear to benefit most from Uber-like work arrangements, and find higher surpluses for demographics such retirement-age workers, women, and some minority groups. We hope that these findings can inform a large current discussion of the benefits a tradeoffs of contractor and gig-work labor laws, and the appropriate types of regulations these jobs should receive. Our first paper in this stream is forthcoming at the Journal of Political Economy.

Current Work: Language Structure as a Driver of Behavior

The idea that language can impact the way people think and act has a rich history in economics, linguistics, philosophy, and psychology and has been a topic of considerable controversy. Saussure, the founder of both structural linguistics and semiotics, characterizes reality as an inherently continuous phenomena that is discretized and organized by language, writing: “if words stood for pre-existing entities they would all have exact equivalents in meaning from one language to the next, but this is not true” (Saussure 1916). Saussure points out that while the concept of “red” can be said to exist, it discretizes an inherently continuous phenomena (wavelength), and is created and organized by language. More recently, the idea that language can influence thought has become known as the Sapir-Whorf hypothesis (SWH, Whorf 1956), and Brown (1976) first

enumerates what has become known as the weak SWH, which claims that differences in linguistic categorization can systematically affect cognition. This hypothesis has generated several interesting lines of research in cognitive linguistics and psychology, which have found robust effects across a number of cognitive domains.

Experimental work has established the validity of Saussurian effects of language on color perception by examining differences between speakers of languages with different color partitions. For example, Russian marks an obligatory distinction between light blue (goluboy) and dark blue (siniy) which English does not; and empirically Russian speakers treat light and dark blue as conceptually distant as blue and green are to an English speaker. Exploiting this variation, Winawer et al. (2007) finds that Russian speakers do better than English speakers in distinguishing blues when the two colors span the goluboy /siniy border (but not when then do not), and these differences are eliminated when subjects must simultaneously perform a verbal (but not a spatial) distractor task. Further implicating language in this differential precision, Franklin et al. (2008) finds that this difference holds for adults, but not for pre-linguistic infants. What this (and several other examples in the psychology literature) teach us, is that obligatory linguistic categories can affect people's ability to categorize, encode, and recall subtle distinctions in the world around them.

As an economist, I became interested in whether these sorts of obligatory distinctions could affect something more closely related to decision making. As a behavioral economist, I focused on an aspect of language, time perception, which connects to the large literature in behavioral economics on discounting.

English and German evolved from the same proto-language (English is a Germanic language). Linguistically, however, there are fundamental differences between the two languages on time perception. When you're talking about a future event in German, it is almost always permissible to speak in the present tense, as long as the context prevents confusing it with present time. The same is true for Chinese, Japanese, and all of the Nordic languages. What was striking to me as an economist is that all of these exemplars of languages with very weak grammatical distinctions between the present and the future are also countries which have extraordinarily high savings rates; rates difficult to explain by purely economic factors. Now, of course this observations isn't really proof of anything: but it started this research agenda. The hypothesis this led me to is that languages which do not grammatically distinguish between present and future events (what linguists call weak-FTR languages) lead their speakers to take more future-oriented actions. Intuitively, fewer distinctions between present and future events puts future benefits more on par with present costs, making it easier to save.

In my first paper on this subject, I show how this prediction arises naturally when well-documented effects of language on cognition are merged with models of decision making over time. Then, I examine whether this hypothesis manifests itself in people's real-world economic decisions. To do this, I adopt the most detailed typological classification I could find in Linguistics, an

examination of cross-linguistic future-time reference strategies produced by the EUROTYP project (a large European Science Foundation funded language typology involving hundreds of linguists from 1990-94). I find a strong correlation between their measure of how a language treats future-time reference (FTR), and the choices that speakers of those languages make when thinking about the future. Specifically, in large data sets that survey families across hundreds of countries, I find a strong and robust correlation between the obligatory marking of FTR in the language a family speaks, and a whole host of forward-looking behaviors, like saving, exercising, and not smoking. What's remarkable is that these correlations hold both across countries and within countries, even when comparing effectively identical families born and living in near each other.

The empirical work revolves around trying to eliminate sources of possible common causation. Obviously, savings behavior of current families can't have CAUSED their language structure. What we have to worry about is the possibility of something like correlated-cultural adoption: the German language spreading along with German values towards savings. More than this, we might need to worry about a whole host of other possible factors: might some languages have co-diffused (spread together) with attitudes towards education? With attitudes towards work? With a particular religion?

Most of my early work explored exactly these types of concerns. For example, if we thought languages co-diffused with proclivities towards education, one way to see that would be to compare only families with identical levels of education. If we though the co-diffusion was mainly spatial, we might think to compare families born and living in the same country, or even city. Effectively, by trying to control for confounding factors that could be driving a spurious correlation between language and behavior, we can try and get a sense of what these patterns across families really mean. This leads to a kind of statistical analysis epidemiologists call conditional-logistic regression. What I do, in effect, is drop families around the world into one of 1.4 billion buckets, where two families fall into the same bucket if and only if they are identical in country of birth and residence, age, sex, income, family structure, number of children, and religion, where the religions of the world are broken up into 74 types. What I then look at is pairs of families who fall into the same bucket, but who report speaking different languages at home.

Now, at this level of detail, even very large economic data sets only allow me to look at around twenty-five-thousand families around the world, who live in nine countries with enough native linguistic diversity to form such pairs. Those countries are Belgium, Burkina Faso, the Dem. Rep. of the Congo, Estonia, Ethiopia, Malaysia, Nigeria, Singapore, and Switzerland. What I find is that in every one of those countries, the categorization of languages I adopt from the EUROTYP project (a large European Science Foundation funded language typology involving hundreds of linguists) seems to have a large and consistent correlation with savings behavior. That is, the direction and the magnitude of these effects are statically identical in every country, despite spanning different regions of the world, and very different languages.

I also investigate a few other ways these effects might be spurious, and find no evidence of any alternative hypotheses. These effects can be measured not only in savings behavior, but in many different future-regarding behaviors studied by economists, smoking, exercise, obesity, and condom use. I find exactly the same effect in all three of those behaviors, in the predicted directions, comparing sets of nearly identical families. The OECD has collected savings data going back through the 70's, and when I look at these effects over time, I find stable effects in every decade from the 70s to the present. The effect also doesn't seem to be driven by minority languages: weak-FTR speakers save more regardless if they are the majority or the minority in their country. In short, the data simply don't seem to suggest that confounding factors play a significant role in explaining the consistent set of correlations between the strength of future reference in language and future-regarding behavior.

The first paper I wrote about these effects was published in the *American Economic Review* and has been well-cited: according to Google Scholar it is in the AER's fifteen most cited paper from 2013. I continue to work in this area, and I hope that several follow up papers to this work will more firmly causally establish the cognitive channels through which this effect occurs.

Most recently, with several linguist co-authors I have begun to develop new statistical techniques to test how robust the effects I find between language structure and behavior, are to the potential co-evolution of language structure and deep cultural parameters; an issue which threatens the causal interpretation of my findings. In an initial paper, we found that the association between language structure and savings behavior is surprisingly robust to statistically tests that measure the typical co-evolution of linguistic features across long spans of time, which is presumably an upper bound on the co-evolution of linguistic features and non-linguistic cultural parameters. In another project, I've been examining direct discounting measures collected as part of the Global Preference Survey, which serves as a more direct test of the linguistic hypothesis. All of the correlations I found in my earlier work with savings behaviors appear even stronger and more robust to phylogenetic tests in the GPS data.

We have also begun studies that attempt more precisely separated correlations, that show for example that future tense structure is much more closely correlated with the measured betta component rather than the delta component of discount functions, a central (but previously untested) prediction of my original paper. I'm excited about this research agenda.

OLDER WORK THAT I AM NO LONGER ACTIVELY WORKING ON

Early Work with Monkeys

My earliest, most numerous, and most cited papers involve experimental work with monkeys. I began this work at Harvard (working with tamarin monkeys), and continued it at Yale, where I worked with capuchin monkeys. When I started this work, though experimental methods with human subjects were already quite common for economists, work on animals was unusual.

A few economists had studied animal behavior before, most notably Kagel, Battalio & Green who studied a variety of economic decisions (e.g., consumer demand, labor supply, risk aversion, and intertemporal choice) in rats and pigeons. Similar to the point made by Gary Becker in his 1962 paper "Irrational Behavior and Economic Theory", KB&G's papers demonstrated experimentally that the central insights of price theory (such as downward sloping demand curves) work well even in the very low-rationality environments of animal experiments. In a parallel literature, psychologists have studied monkeys to examine the evolutionary origins of cognitive systems such as visual processing, numeracy (monkeys can count to about 4), working memory, other-regarding preferences, and even theory of mind. My work on capuchin monkeys (with co-author Laurie Santos) links these two literatures by using a behavioral theory of decision making (prospect theory) to examine the decisions of monkeys which are closely related to humans.

Specifically, we present monkeys with decisions similar to those in which humans display cognitive biases, biases which have been implicated in systematic departures from optimal behavior in human decision making. By looking at how closely-related primates make these same decisions, we examined the evolutionary continuity or discontinuity of the psychological processes underlying the biases humans display. That is, if we find that a particular behavioral bias is present not just in humans but in closely related primates, then it is likely that that bias is an innate behavior rather than a learned response to uniquely human experiences such as markets, institutions or culture.

To do this, we began by introducing a fiat currency to a colony of capuchin monkeys. This allowed us to present them with a large set of potential trades and purchase opportunities. A monkey would first be introduced to small metal washers that we were using as coins. At first, monkeys were rewarded for merely playing with the coins. Later, we introduced them to "traders", human experimenters who would offer a monkey a piece of food in one outstretched hand, but give the food only if the monkey put a coin in the trader's other hand. Different traders (who wore different color uniforms) each traded for different types of food. Our monkeys quickly began to recognize these people: an apple human, a pineapple human, and a grape human. Gradually, our monkeys displayed many signs that they understood the value of these coins: a monkey who found a coin would immediately rush to the trader who sells their favorite food. Our monkeys also quickly learned to recognize and respond to changes in price, buying more when food was on sale and less

when it was expensive. Our monkeys also began to understand that other monkeys also valued these coins: trading coins with each other emerged spontaneously.

Overall, training took about six months, after which almost all of our monkeys were making basic budgeting and purchasing decisions. Once they seemed comfortable with prices and trading, we would introduce a monkey to even more complex trades involving risk. Several new traders acted similarly to the old ones except for one key difference: these traders would sometime add or subtract food before handing it over. For example, one trader would always show a monkey two pieces of apple, then if given a coin, would either take a piece away and only hand over one piece, or add an extra piece and hand over three (each with 50% probability).

With a bit of experience our monkeys seemed to understand gambles, and this new ability allowed us to present them with choices designed to be similar to those used to study decision-making biases in humans. In our first set of experiments, we looked at gambles in which humans typically display both reference-dependence (outcomes are evaluated as gains and losses, not as levels) and loss-aversion (losses motivate more than comparably sized gains). We found that capuchin monkeys display the same prospect-theoretic biases, both qualitatively and quantitatively; monkey and human estimates look identical. Specifically, capuchin monkeys displayed a coefficient of loss aversion (how much more losses feel bad than gains feel good) of around 2.5, almost exactly the mean of human estimates.

The paper we wrote about these first experiments was published in the *Journal of Political Economy* and has been well-cited: according to Google Scholar it is in the JPE's ten most cited paper from 2006. My co-authors and have I extended our earlier work to other components of prospect theory, allowing us to show that the theory as a whole appears to hold broadly in closely-related monkeys.

Our most important follow up paper shows that monkeys, like people, systematically switch from risk-adverse to risk-seeking behavior, when identical gambles switch from being described as gains, to being described as losses. This flip in risk-preference is called the reflection effect, and is one of the three principle components of prospect theory. The experiment we run is simple to describe: in two different experiments, monkeys have to decide between two pieces of apple for sure (option A), or a 50-50 gamble between one and three pieces (option B). What differs between experiments is what a monkey starts with: monkeys begin with either one or three pieces of apple.

If they start with one piece of apple, option A (two for sure) is implemented by always giving the monkey a gain of one piece, while option B entails a 50% chance of gaining two pieces (and a 50% chance of no gain). Conversely, if the monkey starts off with three pieces, Option A is a sure loss of a piece, while Option B is a 50% chance of a loss of two pieces. What we show in our 2010 paper is that monkeys prefer the sure thing when their options are presented as gains, but prefer to gamble when their options are presented as losses. This mirrors the reflection effect in human

behavior, an effect which is thought to be responsible for the disposition effect, one of the most robust anomalies in behavioral finance.

We also wrote a paper testing for the endowment effect, one of the most studied anomalies in behavioral economics. The endowment effect (an experiment replicated hundreds of times in humans), is the pattern that when randomly given either good A or B, the vast majority of people will refuse to trade A for B if given A, but also refuse to trade B for A if given B. That is, people seem to systematically value a good more after they own it, and are reluctant to trade, even if they did not initially receive their preferred object. We show that this also holds with monkeys, even after controlling for trading costs, uncertainty, and waiting times.

Collectively, my monkey research has been cited over eleven-hundred times, and has led many other teams of animal researchers around the world to focus on how behavioral biases do or do not manifest themselves in closely related species.

Early Empirical Work: Crime and Recidivism

In an influential paper on deterrence and prisons, Katz, Levitt and Shustrovich show that very unpleasant prison conditions (specifically, very high in-prison death rates), seem to discourage contemporaneous crime. KLS argued that would-be criminals consider just how bad prison is before committing crimes. This suggested that very harsh and unpleasant prison conditions, though bad for prisoners, might reduce crime rates at large.

My paper with Jesse Shapiro (*American Law and Economics Review,* 2007) asks whether this deterrence effect could be offset by the possibility that harsher prison conditions would make *current* prisoners more likely to re-commit crimes after release. Aggregate numbers suggest this is a large issue: in the Unites States over half a million prisoners are released each year, and over two-thirds will be re-arrested within three years. This alone accounts for a substantial share of crime. Yet, prior to our work, not only did criminologists not have good studies of a recidivism-effect, but they were widely split on whether the effect was negative or positive.

While we had thought that the most natural prediction was a hardening-effect (where tough prisons *increase* post-prison crime) there was a lot of sociological work that suggested harsher prison conditions *reduce* post-prison crime. One of the dominant theories in the sociological literature on crime was *Specific-Deterrence Theory*, which posited that criminals learn about the severity of prison and the attractiveness of crime from their own experiences: hence terrible prison conditions would scare inmates straight. This theory was backed by careful sociological work. Inmates who were incarcerated in harsher prisons appeared to leave with a significantly worse view of prison, and self-report being much more determined to avoid returning. As economists, we wondered if this intention carried through to post-prison behavior.

Measuring this effect is hard, because prisons are not randomly assigned. To overcome the identification problem, we looked for data that would allow us to explicitly control for the

assignment process of people to different types of prisons. We found a small, federal-prison data repository which, for a little over one-thousand inmates released in 1987, had linked these inmates to a newly formed set of state records, allowing us to connect the federal data to state re-arrest records for each inmate's first three years post-release. Also, the federal prison system had very clear rules for determining prison assignment. Each inmate is given a "security-custody score" based on objective criteria such as age, criminal history and severity of current crime. This score is then used to assign inmate to prisons of different security-levels.

Using this score, in a regression-discontinuity design, we estimated the *causal* impact of prison assignment on behavior, comparing inmates who found themselves just on either side of a cutoff sending them to different types of prisons. The regression-discontinuity (RD) design identifies off of the fact that the treatment (here, what type of prison you go to) changes *discontinuously* in your security-custody score, while other characteristics of the inmate are changing continuously.

We find very large effects in this paper: each additional prison security level causally increases the 3 year re-arrest likelihood by more than 10 percentage points. It also has a dramatic effect on the severity of crimes: harsh prison conditions make criminals more likely to recidivate into violent crimes, not profit-driven (pecuniary) crimes.

We published this paper in the ALER, following the Katz, Levitt & Shustorovich paper. Our paper won the American Law and Economics Association's best-paper prize (for the best paper of the last two years), and according to google scholar is the most cited paper in the ALER since 2007, the year it was published. Subsequent researchers have used our technique on richer data sets that allow for more precision, and in total these papers have now contributed a tremendous amount to our understanding of how incarceration affects many aspects of ex-inmates post prison lives and behavior.

Early Work on Cognitive Dissonance: the Effect of Choice on Preferences

The central methodological assumption of micro-economic theories of behavior is revealed-preference: economic actors reveal a coherent set preferences with the choices make. Interestingly, one of the most important theoretical paradigms in social psychology flips this logic on its head. Developed by Leon Festinger in 1957, the theory of cognitive dissonance posits that when two or more cognitions are inconsistent with one another, an uncomfortable state of 'dissonance' is produced. People are motivated to resolve this by changing their cognitions, and because knowledge about one's recent behavior is especially resistant to change, the study of dissonance reduction has typically focused on attitudes that shift in the direction of recent behavior. In other words, dissonance theory posits that behaviors shape preferences, rather than just reflect them.

Since 1957, there have been hundreds of studies empirically testing the specific theory that cognitive dissonance drives people, after making a simple choice between two options, to increase their liking of the option they choose, and devalue the option they reject. My work shows that all of the standard methods of testing this theory, however, are flawed, and yield false positive rates

approaching 1. In later work, Jane Risen and I develop and test a research methodology that correctly tests for this form of "mere-choice induced" dissonance.

Dissonance theory has received support from the results of three experimental paradigms. Developed by one of Festinger's earliest students (Brehm, 1956), experiments using the "free-choice paradigm" find that after making a difficult choice between two items, participants' rating of their chosen alternative tend to rise, and the rating of rejected alternative tend to fall. This phenomena has come to be called 'spreading of alternatives', and was the earliest and most reliable demonstrations of dissonance reduction. Another set of experiments used the "induced-compliance paradigm". Festinger and Carlsmith (1959) asked participants who had just finished a boring task to tell another student that the task was interesting. The researchers found that participants who were paid $20 to lie (high justification) did not change their attitudes because they could easily explain the inconsistency between their action and their attitude. However, participants who were paid only $1 to lie (low justification) expressed a more positive attitude toward the boring task. Finally, using an "effort-justification paradigm", Aronson and Mills (1959) and Gerard and Mathewson (1966) found that participants who had to go through a severe initiation to join (what turned out to be) a pretty dull group liked the group more than those who went through only a mild initiation.

Early dissonance theory researchers hypothesized that attitude change occurred because people were motivated to undo the unpleasant state that was caused by the inconsistency between their attitude and their recent behavior. Thus, the interpretation was that to reduce the experience of dissonance, participants come to like chosen objects more, form a more positive attitude toward boring task, and come to see dull groups as more interesting.

More recent work in social psychology, using children and monkeys as subjects (see Egan, Santos, and Bloom *Psychological Science, 2007*) argue that cognitive dissonance is actually a much more basic process: it occurred just as reliably in animals that don't display a well-developed sense of self (capuchin monkeys), and in amnesiacs who couldn't remember the choices they had made which were the source of purported inconsistencies. These radical new findings suggested that dissonance processes didn't require conscious knowledge of inconsistent cognitions, or even really, cognitions. All of these new papers used the free-choice paradigm, which had become the most widespread experimental test of cognitive dissonance. I will use the set up in the ES&B paper to provide the intuition behind my critique of the empirical work in this area.

ES&B's main experiment begins with children rating a number of objects on a discrete five-point scale. After rating enough objects, three objects that were rated equally (say rated 4) are chosen for use in a second stage of the experiment. Call these three items A, B, and C. The experiment then asks children to choose between a randomly chosen two of these items, say A and B. Calling the object which the child chooses A, the child is then asked to choose between B (the initially rejected item), and C (a random third item that was also initially rated 4). The hypothesis here is that the act of choosing between A and B will *cause* the child to devalue the rejected item B. ES&B

look for this by examining how children then choose between B and C. If a child is more likely to choose C than B in this choice, they are said to have engaged in dissonance reduction. I argued that this was to be expected, and that in fact, a *perfectly* rational person should be *expected* to choose good C 66% of the time. To see this, consider every possible (strict) way people could feel about the three goods A, B, and C, listed in the table below.

Table: All Possible Preference Orderings

	Case 1:	Case 2:	Case 3:	Case 4:	Case 5:	Case 6:
Best:	A	A	B	B	C	C
Middle:	B	C	A	C	A	B
Worst:	C	B	C	A	B	A
Cases where A is preferred to B are shaded. In these cases:						
B vs. C:	B	C			C	

Note that if every possible preference profile is equally likely (which must be true if A, B, and C are randomly ordered), then after a subject has revealed themselves to prefer A over B (by choosing A over B), then we should in fact predict that they would choose C over B two-thirds of the time. This is just a consequence of liking A more than B telling us a bit both about a person feels about A, and on average, how they feel about B. In ES&B, children choose C 63% of the time and capuchins choose C 60% of the time. ES&B argue that since both of these numbers are significantly greater than 50%, both monkeys and young children experience dissonance. I argued that both of these results should be compared to 2/3rds, and that only numbers significantly higher than 2/3rds would demonstrate cognitive dissonance (both number are in fact, statistically indistinguishable from 2/3rds).

Note the fundamental problem here: this experimental design assumed that three goods that were all initially lumped in the "4" bucket must be completely identical, and that people's choices between them could be thought of as random (good A is no different than good B, even though it was chosen). But this is only true if asking a child to sort stickers into five discrete buckets is enough to completely exhaust how they feel about stickers (and false if one of the stickers is actually a 4.1, one 4.26, and one a 4.9). Indeed absent this strong assumption of equality, the belief that C and B should be equally chosen in the third round is mathematically equivalent to a well-known logical fallacy, popularly known as the three-door (or Monty-Hall) problem. Fundamentally, even though subjects are asked to make choices between randomly chosen options, *which good they choose* is not random, and ignoring that may introduce powerful biases into an experiment.

At the most fundamental level, this critique focuses attention on one of the most fundamental principles of microeconomics**:** choices reveal preferences.

Now while ES&B is relatively novel in using a three-goods version of the FCP, what I discovered was that all FCP paper suffered from the same fundamental problem, and that this could be easily quantified. I found that this methodological flaw could explain EVERY pattern in this literature, including the finding in children, amnesiacs, and the interactions researchers had been finding between dissonance with culture and self-affirmation. Thinking that the best place for this work to have impact was in Psychology journals, I began collaborating with psychologist Jane Risen, and together, we did two things. First, I re-wrote the proof to be much more detailed, with separate sections for each sub-type of free-choice dissonance experiments, rather than tying the proof to the information structure common to every FCP experiment. Second, Risen and I ran a series of free-choice experiments which demonstrated that (at least in the logic of this literature), cognitive dissonance was able to make causality run backwards in time. That is, we ran experiments that showed (as predicted by my theorem), that if dissonance was measured the way the literature had been measuring it, then choices that a subject is going to make *tomorrow* (but does not know about) causes dissonance-driven preference change in how they feel *today*. Under certain conditions, these experimental methods also provide a control condition that can be used to isolate the bias we identify, and correctly measure choice-driven dissonance. This became our main paper on this topic, and is published in the Journal of Personality and Social Psychology.

Risen and I followed up on this early paper with several that show that this problem holds even more broadly than we had originally claimed, and offer experimental techniques that avoid the revealed preference problem. By and large, I think that this work has slowly but surely had an positive impact on this important literature. Before our paper over one thousand papers had been published either using or citing the experimental techniques we found to be spurious; now we have seen no paper published with those techniques within the last five years, and many new papers explore alternative methods of measuring choice-induced effects in ways that do not suffer from the problems I identified.