Are Coarse Ratings Fine? Application to Crashworthiness Ratings[*]

Siqi Liu[†], Bhoomija Ranjan[‡], and Benjamin Reed Shiller[†]

[†]Brandeis University

[‡]Monash University

November 2020

Abstract

Many rating organizations intentionally coarsen ratings before public presentation, for example by using a discrete badge rather than a continuous rating. We investigate the impact of coarsening empirically in the context of automobile crashworthiness ratings. Specifically, we construct a univariate continuous crashworthiness rating from crash test measurements and observed fatality rates. We then estimate a random coefficient model of vehicle demand under status quo coarse ratings and simulate outcomes under counterfactual continuous ratings. We find that consumers alter vehicle choices, thereby reducing fatalities by 7.4%, which implies 1,850 fewer U.S. fatalities annually. Finally, we explore whether incentives to produce crashworthy vehicles are reduced enough to offset benefits of finer information. We conclude that a continuous rating format would reduce fatalities.

[*] We would like to thank Ariel Pakes, Joel Waldfogel, Imke Reimers, and colleagues at Brandeis for very helpful comments. We would also like to thank Gary Feld and Leo Yon for generously helping us acquire needed data for this project, and Aiden Zhang for excellent research assistance. Lastly, we would like to thank import.io for providing a discounted rate for their software.

1 Introduction

In addition to gathering, verifying, and providing relevant information to consumers, certification and rating organizations choose how to package and present information. Many organizations intentionally coarsen information before public presentation. For example, the two U.S. organizations that evaluate crashworthiness (safety) of vehicles employ a small number of discrete ratings. A natural question is whether intentionally coarsening ratings further promotes their missions to reduce deaths, injuries, and economic losses from traffic accidents. We investigate whether a continuous crashworthiness rating would reduce fatalities relative to a counterfactual scenario with discrete ratings.

It is an empirical question whether coarsening quality ratings improves outcomes. *Ceteris paribus*, continuous ratings provide finer information to consumers. However, there are several reasons why organizations may prefer to coarsen ratings. For example, coarsening or otherwise repackaging ratings might increase salience (Dai and Luca, 2020; Luca and Smith, 2013; Pope, 2009). However, others find that coarsened certifications may crowd out consumer efforts to process more precise information, consequently worsening outcomes (Farrell et al., 2010; Houde, 2018). We abstract away from these issues and instead focus on another mechanism. We argue in Section 2.3 that producers may provide less (more) than the welfare maximizing investments in safety provision when ratings are continuous and marginal consumers have less (more) than average valuations for safety. Coarsening ratings may address these inefficiencies by altering firm incentives, at least in theory (Costrell, 1994; Dubey and Geanakoplos, 2010; Zapechelnyuk, 2020).

Even if coarse ratings are optimal in theory, they may not be in practice. Even with mechanism design expertise, certifiers may be unable to precisely predict how manufacturers will respond to ratings thresholds due to inherent uncertainties.[1] The risk of poorly setting thresholds for coarse ratings may outweigh any theoretical gains from coarsening ratings.

The economics literature has examined a variety of mechanisms policy makers can use to yield desired outcomes. In the context of vehicles, government organizations can require certain features (Golovin, 2019) or tax undesirable features (e.g., the gas guzzlers tax). A less heavy-handed approach is to report quality ratings to consumers. Others find that ratings do influence choices (Anderson and Magruder, 2012; Bollinger et al., 2011; Hastings and Weinstein, 2008; Jin and Leslie, 2003; Jin and Sorensen, 2006; Luca, 2016; Tadelis and Zettelmeyer, 2015), and that firms strategically manipulate and selectively disclose quality ratings (Fan et al., 2016; Garate and Newbury, 2019; Luca and Smith, 2015; Luca and Zervas, 2016; Mayzlin et al., 2014; Proserpio and Zervas, 2017). For an overview, see Dranove and Jin (2010). We focus on a related topic: the impact of ratings' format choices.

Vehicle crashworthiness ratings are an auspicious context for study. The gains from

[1] In the context of crashworthiness ratings, safety certifiers appear to follow the medical literature—not the economics literature—when designing coarse rating thresholds.

improved ratings may be large; vehicle accidents take a significant economic toll of $242 billion (in 2010) and are the 11th leading cause of death in the United States.[2] If finer ratings were made available, would customers choose different vehicles? Would manufacturers invest more or less in safety? Would there be fewer vehicular fatalities?

To investigate consumer choices under a counterfactual rating scheme, we first construct continuous crashworthiness ratings by relating fatalities to vehicle features and measurements from staged crash tests, yielding vehicle-specific probabilities of driver death. Next, we estimate vehicle demand under status quo coarse crashworthiness ratings using a random coefficient discrete choice model. We then simulate consumer choices and resulting death rates under a continuous crashworthiness rating, and examine the plausibility that a continuous rating could alter incentives for safety provision in a way that offsets gains from providing finer information to consumers.

A challenge when estimating crashworthiness from fatality data is that fatality rates depend on two components: (i) a vehicle's inherent ability to protect its occupants (our object of interest), and (ii) the behaviors of drivers who select different vehicles. The latter has been shown to be potentially large: Levitt and Porter (2001) found that drunk drivers are 13 times more likely to be involved in fatal accidents, and Jin and Vasserman (2018), Reimers and Shiller (2019), and Soleymanian et al. (2019) have shown that drivers have heterogeneous accident risks and decrease accident risk dramatically when explicitly monitored and financially incentivized to avoid unsafe behaviors.

Of particular concern for our study is the possibility that latent driver behaviors and crash test ratings are correlated. For example, if drivers with safe driving habits disproportionately choose vehicles with better reported safety ratings, then we risk attributing the reduction in fatalities from cautious habits to a vehicle's inherent ability to protect its occupants.[3]

Existing studies have attempted to address this selection concern in various ways. Some studies assume away this selection problem (Metzger et al., 2015; Teoh and Lund, 2011). Harless and Hoffer (2007) use vehicle nameplate fixed effects to control for varying habits across drivers of different nameplates, but they discount the possibility that cautious drivers may choose a given nameplate only when it is awarded a safety badge. Ryb et al. (2010) explicitly control for accident speed in expertly reconstructed accidents, but speeds are lumped into two groups, which limits the effectiveness of these controls. Others (Farmer, 2005; Kullgren et al., 2010; Lie and Tingvall, 2002) focus on two-vehicle accidents to control for the magnitude of force at impact (following Newton's third law of motion). However,

[2] https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812013

[3] Safety improvements might directly impact driving behaviors of rational consumers (Peltzman, 1975) by reducing the costs/risk associated with bad driving behaviors. Empirical tests of this *offset hypothesis* yield inconsistent findings (Cohen and Einav, 2003; Peterson et al., 1995; Sen, 2001; Traynor, 1993; Winston et al., 2006).

it remains possible that the more aggressive/faster vehicle may experience worse secondary outcomes after the initial collision or safety features like well-designed crumple zones limit injury risk of occupants in both cars, including the vehicle collided with, thereby confounding comparisons. In Section A.3, we provide evidence that the selection of safer drivers remains an important confounder in the two-vehicle crash analyses used in the best previous studies of vehicle crashworthiness. Our method addresses the selection issues that confound results in the existing literature and estimates overall crashworthiness across the empirical distribution of different accident types, including single-vehicle accidents (e.g., vehicle crashing into a tree), which accounted for 55% of fatal accidents in 2017.[4]

Our method for addressing selection of drivers with safer habits is similar to the regression discontinuity strategy used to estimate the causal impact of ratings on consumer choices (Luca, 2016; Anderson and Magruder, 2012). We include both continuous measures of crashworthiness (which are not observed by consumers), as well as observable discrete crashworthiness measures as explanatory variables. Fatality risk presumably increases continuously in measures of injury risk and cabin intrusion. But a discontinuity might exist at the threshold for a higher reported discrete rating if driver selection to safer-rated vehicles is meaningful. In Luca (2016) and Anderson and Magruder (2012), the discrete change at the discontinuity is the object of interest. However, our goal is to estimate the discontinuity only so we can remove it (and thus the impact of driver selection) from our fatality rate predictions, thereby inferring the causal relationship between these measures and the likelihood a typical driver dies while driving a given vehicle.

Our estimates show that continuous ratings provide substantially more information, leading consumers to alter vehicle purchase decisions in counterfactual simulations. As a result, simulations suggest that switching to a continuous rating would reduce fatalities by 7.4%, implying an annual reduction of 1,850 fatalities in the United States. Consumer welfare is predicted to rise by about $7 billion over five observed years. It is possible that manufacturers would respond to continuous ratings by lowering investments in crashworthiness. However, we provide evidence that manufacturer response would need to be large to offset these gains, and we document market features that suggest the benefits of a discrete rating are small in this setting.

The rest of the paper is organized as follows. Section 2 provides an industry background and discusses the theoretical benefits of coarse ratings. Section 3 describes the data. Section 4 describes our continuous crashworthiness ratings, and Section 5 presents the model of consumer demand for vehicles. Section 6 reports demand estimates and counterfactual predictions. A brief conclusion follows.

[4] Jacobsen (2013) uses fatalities in single-car accidents beyond levels implied by crash tests measurements to control for relative differences in driving behaviors across vehicle classes. However, to separate signal from substantial inherent noise in vehicle level fatality rates due to their rarity, vehicles are pooled into ten groups, obscuring the differences between vehicles in the same class required for our analyses.

2 Background

2.1 Organizations That Rate Crashworthiness

In the United States, there are two well-known institutions offering vehicle crashworthiness ratings: the National Highway Traffic Safety Administration (NHTSA), a government agency, and the Insurance Institute for Highway Safety (IIHS), an independent non-profit historically funded by insurers. The stated mission of both is to improve automotive safety.

A major and growing focus of both organizations is crash test safety ratings. The NHTSA and IIHS both began reporting crashworthiness ratings in the 1990s.[5] Initially, each used just one crash test: a frontal collision into a fixed barrier.[6] The NHTSA and IIHS added side crash tests in 1996 and 2003, respectively.[7] The NHTSA added a measure of rollover risk in 2000, and the IIHS added a roof strength test in 2009. In 2012, the IIHS added a more stringent frontal crash test where only 25% of the front collides with the barrier (on the driver's side), and in 2017 they added an analogous test for the passenger's side. Both organizations note that comparisons of frontal collision tests are valid only between vehicles of similar weight, whereas side and rollover tests are comparable across weight classes. The IIHS has recently added tests evaluating headlights and automatic emergency braking, although these ratings were typically based on optional packages or select trims.

Crashworthiness ratings are highly visible. The ratings from the IIHS and NHTSA are prominently shown in automobile reviews on websites like ConsumerReports.com, Edmunds.com, and USNews.com.

2.2 Format of Ratings

The continuous measures from crash tests are reported in inconspicuous technical reports, and understanding many continuous variables may be challenging and time-consuming for consumers. Therefore, both agencies transform the continuous measurements into simplified scores that are featured on their respective websites and publications. Both agencies elect to report discrete scores.

The NHTSA reports separate discrete scores for each crash test type (e.g., side, front) on a five-star scale, and report an overall rating also on a five-star scale. Most vehicles earn four or five stars. Ratings below three stars are rare.

The IIHS reports a discrete four-point scale for each test type (good, acceptable, marginal,

[5] Crash test milestones: NHTSA: https://www.nhtsa.gov/ratings, IIHS: http://www.iihs.org/iihs/about-us/milestones.

[6] The NHTSA uses a full head-on collision, and the IIHS uses a partial overlap collision.

[7] The NHTSA side-crash test involves a car-like object crashing into the tested vehicle, and the IIHS involves a SUV/truck like object with a higher point of impact on the tested vehicle.

or poor), and then awards "Top Safety Picks" based on evolving criteria. A link to vehicles awarded "Top Safety Pick" badges is prominently shown on the ratings landing page (see Figure 1), and "Top Safety Pick" badges are also prominently shown on pages for specific vehicles.

The relationship between discrete and continuous scores is well illustrated by the IIHS's side-impact crash test. A separate discrete subrating (good, acceptable, marginal, or poor) is assigned to the vehicle structure and each of four separate injury regions. The scores from these five subcategories are combined to reach the overall side-impact rating, using a demerits-based system described shortly.

The vehicle structure subrating in the side-impact crash test is based on a single continuous measurement: the final location of the vehicle's central side pillar (known as the B-pillar) relative to the driver's seat center line, after the stationary vehicle is struck by a 1500 kg (about 3300 lb) barrier moving at 50 kph (about 31 mph).[8] The extent of intrusion is mapped to a discrete subrating according to the criteria in Figure 2.

The injury subratings for each body region (head, neck, torso, pelvis/femur) are intuitively similar. Each injury region includes several distinct measurements, each with its own discrete score. The discrete rating for the injury region equals the worst discrete rating among its subcomponents.[9]

Finally, the five subratings (vehicle structure, as well as injury regions: head, neck, torso, pelvis/femur) are combined into an overall side-impact crash test rating using a demerits-based system depicted in Appendix Table A2.[10] Ratings for other crash tests (e.g., moderate overlap frontal collision) are constructed similarly.[11] The criteria for awarding "Top Safety Pick" badges, which are prominently shown, change most years, but are based on the discrete scores for each crash test type.

The NHTSA's mapping of continuous measurements to discrete scores (stars) is less complicated. Before 2011, only two injury measures were used in the frontal crash test, the Head Injury Criteria (HIC), and the chest G-force. Awarded number of stars depended on total chance of serious injury to either the head of chest (see Appendix Figure A1). The NHTSA's discrete crash rating for side collisions was calculated similarly. In 2011, to make the ratings more stringent, the NHTSA included additional injury measures, used more stringent thresholds along with a smaller crash test dummy, and added a side-pole crash

[8] "IIHS Side Impact Test Program – Rating Guidelines." https://www.iihs.org/media/2104caa9-7f7e-41fa-a4a5-af65f1cab89e/l9AWAw/Ratings/Protocols/current/side_impact_guide.pdf

[9] "Side Impact Crashworthiness Evaluation – Guidelines for Rating Injury Measures." https://www.iihs.org/media/ba19c647-c2e7-4341-8f12-fe84b0e68a21/9mhzGw/Ratings/Protocols/current/measures_side.pdf

[10] "Side Impact Crashworthiness Evaluation – Weighting Principles for Vehicle Ratings." https://www.iihs.org/media/eda4bd5a-06a8-4c8e-9b7e-a89ce7b822b0/snHVbw/Ratings/Protocols/current/side_impact_weighting.pdf

[11] IIHS thresholds for continuous measurements have not changed over time.

test.[12]

2.3 Ratings Format and Welfare

In this subsection, we first argue that firms may provide more or less than the utilitarian welfare maximizing level of investment in safety features when the exact safety level is conveyed to consumers via a continuous crashworthiness rating. We then argue that a discrete rating system may alleviate these inefficiencies in theory, but may not in practice.

When choosing safety levels of their vehicles, firms trade off the costs of improving crashworthiness with the corresponding increase in revenues. The extent to which firm j's revenues increase as crashworthiness of their vehicle improves depends on the value placed on safety among consumers who are nearly indifferent between firm j's product and some other firm's product. Firms will invest more in safety if these marginal consumers have higher valuations for safety, and vice versa. How much captive consumers value safety may be irrelevant; small changes in crashworthiness may not impact their product choice.

However, a social planner should trade off the costs of improving safety with the total benefit to all consumers and third parties, including consumers who are captive to one of the firms. Assume for the moment that the only two parties impacted by investments in safety are consumers and manufacturers. If marginal consumers have much higher valuations for safety than captive consumers, then firms may provide more than the utilitarian welfare maximizing level of investment in safety features. Alternatively, if marginal consumers have relatively low valuations for safety, firms may provide less than the utilitarian welfare maximizing levels of investments in safety features. Thus, perfect information conveyed through continuous crashworthiness ratings does not necessarily result in firms providing welfare maximizing levels of safety features.

Additionally, a social planner may consider the welfare of third parties such as insurers, first responders, hospitals, and relatives and friends of those harmed in accidents. These additional considerations may also influence the utilitarian welfare maximizing level of safety provision.

Suppose a regulator would prefer that firms invest less in safety provision than they do when a continuous safety rating is reported. A regulator could achieve this goal by reporting a binary measure of whether the welfare maximizing level of safety has been met or exceeded. Firms would not be recognized for safety improvements beyond the threshold and thus would have no incentive to provide them.

Perhaps a more interesting question is whether a more stringent binary safety threshold could induce firms to invest more in safety features compared to when a continuous measure is reported. Firms whose effort under continuous ratings would be near but still below

[12] https://www.consumerreports.org/cro/2011/08/crash-test-101/index.htm

the threshold after coarsening have a choice. If they increase their effort slightly, they will be recognized as safe after ratings are coarsened. Otherwise, after coarsening they will be pooled with unsafe vehicles and recognized as unsafe. Costrell (1994) and Zapechelnyuk (2020) show that such agents below but near the threshold increase effort when ratings are coarsened. Thus, coarsening can increase incentives, at least for some agents. However, agents lack incentives to exert effort beyond the coarse ratings threshold because their additional effort would not be observed. Therefore, other agents who already exceeded the threshold may reduce their effort to the threshold after ratings are coarsened. One might mitigate this concern by adding additional discrete ratings for such agents to strive for. But using more thresholds diminishes the incentive effects from having discrete thresholds (Dubey and Geanakoplos, 2010).

While discretizing ratings may improve outcomes in theory, it may not in practice because poorly set thresholds may reduce welfare. With a discrete rating system, firms are not recognized for exerting effort beyond the threshold and thus have no incentive to do so. If the threshold is set too low, then firms will underinvest in safety features. If the threshold is set too high, then firms may forgo attempts to meet the threshold. If there is inherent uncertainty as to how firms will respond, then there may be no threshold that improves ex-ante expected investments in safety. The risks of setting a harmful threshold may outweigh the small gains achieved if the ex-post optimal threshold happens to be chosen.

3 Data

We use a novel dataset of automobile sales, characteristics, and fatalities compiled from multiple data sources. Because the component datasets have varying units of analysis and time horizons, we construct two main datasets: one used for demand estimation, and the other for analyzing vehicle crashworthiness. We describe both datasets in turn below.

3.1 Data for Demand Estimation

The first set of data are used to estimate demand for vehicles. We use monthly new automobile sales in the United States by nameplate and model year provided by the Alliance of Automobile Manufacturers. The data span 60 months, from January 2013 to December 2017, and include all vehicles for sale during that period, including model years 2012 and 2018.[13] We exclude passenger vans and “supercars” with real prices exceeding \$50,000 measured in 1983 dollars (about \$130,000 in current dollars).

We link these data with vehicle characteristics sourced from Wards Automotive Reports.

[13] We limit observations of a vehicle to the first 24 months it is for sale, and censor monthly sales below 25 units.

These include static characteristics such as car dimensions (length, width, height, weight), body style, drive type (2WD/4WD), horsepower, combined fuel economy, and manufacturer suggested retail price (MSRP).[14] We also link these data with U.S. Environmental Protection Agency (EPA) data on which vehicles were subject to a "gas-guzzler" tax and which hybrid and electric vehicles were eligible for a federal tax credit.[15] We also note vehicle class designated by Wikipedia: luxury, sports, executive, or crossover.

To these data, we add safety ratings information recorded by the IIHS. For each nameplate and model year combination evaluated by the IIHS, we record the continuous measurements from staged crash tests and the discrete ratings reported to consumers (see Section 2.2 above). IIHS also records whether the vehicle was awarded either a "Top Safety Pick" or "Top Safety Pick+" badge, the latter of which typically depended on optional features not included on the base model.[16] We subsequently designate any vehicle that receives an IIHS badge of either type as an IIHS badge award winner.

We combine these time-invariant data with time-varying customer incentives (price discounts) by nameplate/model year combination, from Automotive News' Data Center. These rebates provide a useful source of variation in monthly prices for the vehicles in our data. Our final price variable equals MSRP of the base model less any consumer incentives and federal tax credits for qualifying hybrid/electric vehicles, plus the "gas-guzzler" tax if applicable. Price is measured in 10,000s of 1983 dollars, using the consumer price index to deflate. Additionally, we use monthly petroleum prices and vehicle fuel economy to construct time- and vehicle-specific gas mileage per dollar (MPD), measured in hundreds of 1983 dollars.

Finally, we collect information related to the number and types of households in the market. We assume each household buys at most one new vehicle each year. The assumed monthly market size equals the total number of households in the United States divided by the (12) months in the year.

The final dataset records monthly sales and prices of 1,727 nameplate/model-year combinations from 2013 through 2017. Trends in vehicle sales and characteristics are summarized in Table 1. Note that the fraction of vehicles awarded an IIHS badge fluctuated substantially, in part due to evolving standards.[17] However, a typical consumer may have been unaware of these evolving criteria because the changes were not prominently noted. Average monthly sales for 2012 and 2018 model years are lower than other model years because these vehicles were in the waning or waxing end of their market evolution cycle. The strong non-monotonic relationship between a vehicle's age and sales, depicted in Appendix Figure A2, suggests that we must be cognizant of the impact of age on novelty and inventory variety—of colors,

[14] In case of multiple trims, we use characteristics for the base model for that year. Any missing car characteristics are manually imputed from Edmunds.com.

[15] `see:www.fueleconomy.gov` and `www.epa.gov/fueleconomy/gas-guzzler-tax`

[16] Vehicle nameplate model-years that are not rated by IIHS but do record sales are designated as not receiving a "Top Safety Pick" badge.

[17] See, for example: `https://www.autoblog.com/2013/12/19/iihs-2013-safest-vehicles-video/`.

trims, and optional package (Copeland et al., 2011)—when estimating demand.

3.1.1 Microdata

We supplement the aggregate data with individual-level data from the 2017 National Household Travel Survey [NHTS] (U.S. Department of Transportation, 2019). These data are used to construct the distribution of demographics and find associations between household traits and product characteristics to form supplemental micro moments.

The NHTS dataset contains household demographics and vehicle ownership for 129,696 households. With provided weights, the data constitute a representative sample of U.S. households to draw agent data from.[18]

We find meaningful correlations between household traits and product characteristics of chosen vehicles. The correlation between a household's highest education level on a three-point scale [1 = no college degree, 2 = bachelor's degree, 3 = graduate degree] and IIHS badge indicator is $\rho = 0.07$. We also observe a strong correlation between rural indicator and IIHS badge indicator ($\rho = -0.07$). Additionally, we observed strong correlations between vehicle size and indicators for children in household ($\rho = 0.14$) and rural location ($\rho = 0.16$), and between minivan indicator and child indicator ($\rho = 0.18$). The analogous set of covariances are used to form the micro moments described in Section 5.1. The relationship between child indicator and IIHS badge indicator was disregarded because, surprisingly, the pairwise correlation was quite weak ($\rho = -0.01$).

3.2 Data for Constructing Counterfactual Crashworthiness Ratings

The second dataset includes information on the IIHS safety badge status and crash test measurements for 3,011 distinct vehicles (nameplate and model-year combinations) between model years 2005 and 2018 from the IIHS's website. These data contain numerous measures of extent of cabin intrusion and dummy injury severity. Trends in select crash test measurements are reported in Table 2. Note that the measures have declined in magnitude over time, implying lower levels of injury and less cabin intrusion for newer models, at least on average.

These crashworthiness measurements are combined with vehicle characteristics from Wards Automotive Reports, quarterly driver fatalities from the Fatality Analysis Reporting System (U.S. Department of Transportation, c), and cumulative production from the Early Warning Reporting database (U.S. Department of Transportation, b).[19] The result-

[18] Monthly annualized income—measured in inflation-adjusted 1983 dollars—are imputed from static income data and the relative CPI index in other periods. Other demographics are assumed unchanged throughout.

[19] Any missing vehicle characteristics are again imputed manually from Edmunds.com.

ing dataset contains time-invariant vehicle characteristics and crash test measurements separately for each combination of nameplate and model year, along with corresponding quarterly fatalities and cumulative production between 2005 and 2017.[20]

Because driver fatalities are only one among a range of downstream crash-related outcomes, we assemble an analogous dataset on crashes from the Crash Report Sampling System ((U.S. Department of Transportation, a); also known as the General Estimates System pre-2016). These data are collected across 60 sampling sites across the United States for the years 2011-17. Unlike the FARS data wherein records exist *only* for accidents where one or more people have died, these data enumerate outcomes for all crashes reported to police in the sampling sites and measure three possible injury levels: minor, major, or death. We similarly combine these data with information on IIHS safety badge status and crash test measurements, vehicle characteristics, and cumulative production.

3.3 Preliminary Evidence of Manufacturer Response to Ratings

The response of manufacturers to new crash test categories provides strong suggestive evidence that manufacturers design vehicles with crashworthiness ratings in mind. Figure 3 shows the evolution of average IIHS crashworthiness ratings (on a four-point scale from 1 [“poor”] to 4 [“good”]), separately by crash test type and model year. Note that when new tests are introduced, average ratings are rather low but improve quickly, suggesting that manufacturers respond. Note, for example, that in 2012, when the driver’s side small overlap frontal crash test was introduced, the average score was about 2 (marginal). In response, manufacturers redesigned the structure of 97 different models for the U.S. market.[21] By 2016, three fourths subsequently scored “good,” the best possible rating.[22] However, the IIHS remained suspicious that manufacturers were intentionally focusing on safety improvements captured by the driver’s side crash test and were ignoring the untested passenger’s side. After confirming these suspicions using a small sample of vehicles, the IIHS added the passenger’s side small overlap frontal test to their protocol in 2017. As Figure 3 shows, average scores on the passenger’s side were much lower than the scores on the analogous test on the driver’s side in 2017 and 2018, confirming the suspicion that manufacturers had primarily strengthened parts of the vehicle that they expected to be directly tested by independent ratings organizations.[23]

Additionally, there is evidence of clumping of continuous measurements just surpassing the thresholds. We demonstrate this using the IIHS side impact crash test, which was

[20] We use FARS’ vehicle identifiers as the unit of observation in the fatality analysis dataset. Their identifiers sometimes combine multiple nameplates in a single identifier.

[21] `https://www.iihs.org/news/detail/small-overlap-gap-vehicles-with-good-driver-protection-may-leave`

[22] Ibid.

[23] Among vehicles with a passenger’s side small overlap frontal crash test rating in 2017, the average passenger’s side rating on a four-point scale (from 1=poor to 4=good) was 2.3, much lower than the corresponding average driver’s side rating (3.4).

introduced just prior to the beginning of our sample (introduced in 2003), and which records a smaller number of underlying continuous measurements (16 measurements). Recall that each continuous measurement is assigned a discrete score, which are then combined to yield an overall rating for the side crash test. We focus on the four continuous measures that had the lowest average assigned discrete scores in 2005, thus having the most room for improvement. Figure 4 shows histograms of these continuous submeasures, in both 2005 and a decade later. Note that higher values imply more cabin intrusion and greater injury severity; thus, lower scores are better. For two of these measures—the B-pillar measure of structural integrity and rib deflection—there is obvious clumping of scores just below the thresholds, which, if exceeded, yield worse subratings. In 2015, the clumping is more pronounced near the lowest threshold, which corresponds to the highest ratings. We view this clumping as evidence that manufacturers respond to tests, especially considering that there are factors that should mitigate such clumping: (i) there are many submeasures included in the ratings that are simultaneously impacted by vehicle design choices, and (ii) the IIHS is not the only crash test organization (although is arguably the most stringent), because there are other domestic (NHTSA) and international (e.g., Europe's NCAP and Japan's JNCAP) ratings agencies that may also influence manufacturers' design choices.

3.4 Inferring Mechanisms for Improving Safety

A pertinent question is whether safety improvements mostly impact variable costs, or whether they incur higher fixed costs of development. To investigate, we analyze whether firms are more likely to improve safety enough to earn a safety badge when they redesign their vehicle. Specifically, we collapse the data so that each nameplate and model-year combination appears only once, and restrict the data to nameplate/model-year combinations that were tested by the IIHS.[24] We then regress an indicator for newly acquiring a safety badge on an indicator denoting whether the model year was a major redesign, yielding the following results:

$$\begin{array}{ccccccc} 1(\mathit{Newly\ Earned\ Badge}_{v\tau}) & = & \underset{(0.01)}{0.03} & + & \underset{(0.01)}{0.11} \times 1(\mathit{Newly\ Redesigned}_{v\tau}) & + & \epsilon_{v\tau}, \end{array} \tag{1}$$

where v denotes nameplate and τ denotes model year. The results of this linear probability model imply that nameplates that did not experience a major redesign began earning a safety badge only about 3% of the time, whereas newly redesigned vehicles began earning a safety badge $0.03 + 0.11 = 14\%$ of the time.[25] Hence, most observed safety improvements

[24] We exclude nameplate/model years observations when the preceding model year was not evaluated by the IIHS.

[25] Similar results were obtained in a analogous regression that included model year fixed effects to account

coincided with major redesigns, suggesting that many safety improvements require a fixed cost of development. Unfortunately, without access to worldwide sales and profits of vehicles, it is not possible to estimate the fixed costs of improving safety. It is therefore not possible to predict investments in safety provision under counterfactual rating schemes. Instead, we investigate whether market features are favorable for using coarsened ratings to increase incentives for safety provision.

4 A Model of Vehicle Crashworthiness

4.1 Estimating Fatality Risk

In order to evaluate counterfactual scenarios with a univariate continuous crashworthiness rating—which did not exist—we must first construct such a rating. We do so by relating driver deaths in vehicle j and period t to static crash test measurements and characteristics and time-varying controls.[26] We use a binary logit model. Let the probability a driver of vehicle j dies from an accident in period t equal $F(x_{j\ell t}|\gamma_{\ell t}) = \frac{\exp(\sum_\ell x_{j\ell t}\gamma_{\ell t})}{1+\exp(\sum_\ell x_{j\ell t}\gamma_{\ell t})}$, where $x_{j\ell t}$ denotes crash test measurements, vehicle characteristics, and other controls, and $\gamma_{\ell t}$ denotes parameters to be estimated. The log-likelihood objective function is:

$$LL = \sum_j \sum_t D_{jt} \times ln\left(F\left(x_{j\ell t}|\gamma_{\ell t}\right)\right) + \left(Q_{jt} - D_{jt}\right) \times ln\left(1 - F\left(x_{j\ell t}|\gamma_{\ell t}\right)\right), \qquad (2)$$

where D_{jt} denotes the number of drivers of vehicle j dying in period t, and Q_{jt} denotes cumulative production of vehicle j.

There are two main threats to causal inference. First, there are several potential confounders that may be correlated with crash test measurements. The first is vehicle age. If wear and tear reduces crashworthiness, we would observe higher fatality rates for older vehicles.[27] Because older model-years also tend to have worse crash test measurements, we might falsely attribute the higher fatality rates (in later years) to lower safety ratings if we do not control for vehicle age. Similarly, if there are any time-varying reductions in fatalities due to factors other than vehicle design (e.g., improved lighting on highways, drunk driving programs), then vehicles released later face safer road conditions on average. Because vehicles released later also tend to have better crash test measurements, we risk conflating the

for variation in badge requirement stringency.

[26] We focus on driver deaths, as is common in the literature, because every car has a driver. Focusing on passenger deaths conflates crashworthiness with variation across vehicles in the frequency in which passenger seats are occupied.

[27] We remain agnostic to the reason vehicles have more driver deaths as they age. In unreported analyses using public traffic citations data, we find evidence that the rate of traffic citations increases with vehicle age, suggesting changes in driving composition at least partially explains the increased rate of driver deaths as vehicles age.

impact of improved vehicle designs and road conditions if we do not control for time effects. Finally, safety features added in later years that are not measured by crash tests may also be correlated with safety features that are measured by crash tests. Therefore, to yield an unbiased estimate of the influence crash test measurements have on fatality rates, we must control for the year each vehicle was designed. These confounders (age/time/other design trends) are all observed. Whether they can be included as controls depends on whether or not they are separately identified.

A feature of the auto industry can be exploited to control for the aforementioned confounders: age, time/changing road conditions, and design year/trends in unobserved safety improvements. Blonigen et al. (2017) note that changes between model years are typically superficial. Major redesigns occur only about every 5 years. In fact, the IIHS exploits this feature, assigning the same crash test ratings and measurements to all model-years between substantive structural redesigns. For example, the 2010 model-year and 2014 model-year Ford Tauruses are structurally nearly identical and should be approximately equally crashworthy immediately after leaving the production line. Thus, following the intuition from Farmer and Lund (2006, 2015), differences in fatality rates between 2014 model-year vehicles in 2014, and 2010 model-year vehicles in 2010, if not redesigned in the interim, can be attributed to changing road conditions. Differences in fatality rates occurring in 2014, between the 2010 and 2014 model-years, can be attributed to the differences in their age because they face the same road conditions. After controlling for age and time, marginal differences in fatality rates for vehicles redesigned in the interim can be attributed to design changes, which are measured by design-year fixed effects and crash test measurements.

The second threat to causal inference is driver selection. For example, a consumer who is unusually concerned about safety might both choose to drive exceptionally cautiously and to buy a vehicle with good observed crashworthiness ratings. If not addressed, one might conflate the impacts of driver habits and crash tests measurements on driver death rates. In Appendix Section A.3, we provide evidence suggesting that the gold-standard method previously used to control for this selection issue was inadequate.

Our identification strategy exploits the discrete nature of reported ratings. Suppose, for example, that two nearly identical vehicles earn identical ratings on all crash test measures except one. Suppose cabin intrusion in the B-pillar test measure is 12.4 cm in one vehicle, and 12.5 cm in the other vehicle. Because only the latter exceeds the threshold, only the former vehicle receives a top safety score. Hence, while these measurements are very similar, implying nearly identical crashworthiness, consumers may believe that the first vehicle is meaningfully safer. This might lead drivers who care more about safety, and who might drive exceptionally cautiously, to sort into the vehicle with the badge, even though the two vehicles are approximately equally crashworthy. Therefore, differences in fatality rates across these two vehicles are presumably attributable to driver composition rather than vehicle design. Similarly, differences in crash measurements that do not lead to different discrete ratings are presumably unobserved by consumers and therefore should not lead to sorting of cautious drivers. Hence, differences in driver death rates between vehicles with the same discrete rating observed by consumers but different crash test measurements (observed only by the researcher) can be attributed to differences in vehicle design.

Intuitively, this reasoning is analogous to the reasoning behind a regression discontinuity design. However, unlike the standard regression discontinuity design, the break at the discontinuity (for a different safety rating) is not the object of interest, but rather a confounder that must be accounted for and removed from predictions.

To account for the impact of selection of cautious drivers, we include both the discrete score—an indicator variable denoting whether the vehicle was recognized with an IIHS "Top Safety Pick" or "Top Safety Pick+" badge—and continuous measures of injury and cabin intrusion from staged crash tests. Then, when predicting a vehicles' inherent crashworthiness using the model, we reassign the value of the IIHS badge indicator to the production-weighted average value across vehicles, thereby removing impacts of driver composition on predicted fatality rates.

In Table 3, we report our estimates for various specifications. In all specifications, we omit the first seven quarters after a given vehicle (denoted by nameplate and model-year) commences production to avoid biases arising from unknown numbers of unsold vehicles sitting on dealer lots.

The first column of Table 3 includes only age and time fixed effects and the year the vehicle was last redesigned. Next, we consider the 26 continuous measurements of cabin intrusion and dummy injury from the moderate-overlap frontal crash test, the 16 continuous measurements from the side-overlap test, and the discrete safety rating (has safety badge/does not). To alleviate concerns of overfitting, we incorporate and estimate a LASSO penalty parameter using 10-fold cross validation. We apply the penalty only to the continuous crash-test measurements and restrict their coefficients to have the anticipated (weakly positive) sign, implying more intrusion and more force to the crash test dummy raises fatality risk. Estimating an unrestricted estimation model after selecting variables via LASSO regularization has been shown to improve model fit while retaining convergence rates (Belloni et al., 2011, 2013).[28] Thus, we report the unpenalized binary logistic model with the 16 selected variables in Column 2. In the third column, we add vehicle dimensions as additional explanatory variables. In Column 4, we include vehicle weight, horsepower and fixed effects for vehicle body style and class. In the final column, we put in fixed effects for each combination of make, nameplate, and design period (e.g., all model years between major redesigns). These fixed effects approximate all differences in crashworthiness across vehicles.

Next, we compare the relative values of the pseudo R-square across models. Adding

[28] Also see discussion in Section 3.8.5 in Friedman et al. (2001), `https://web.stanford.edu/~hastie/ElemStatLearn/printings/ESLII_print12.pdf`

crash test measurements increases the pseudo R-square from 0.0021 to 0.0062. Adding in size and dimensions increases the pseudo R-square to 0.0081. Including fixed effects for car body style and class further increases the pseudo R-square to 0.0095. Including fixed effects for each combination of nameplate and design generation, thereby accounting for all meaningful design differences across vehicles, only raises the pseudo R-square to 0.0129, due to the inherent randomness of fatal accident frequency for a given nameplate in a given quarter. The pseudo R-square remains low in the last specification, even though the last model is likely overfit. Hence, this exercise suggests that much of the explainable variation in fatality rates across vehicles is attributable to continuous crash test measurements from the front and side crash tests and basic vehicle characteristics.

Note that the coefficient on year designed is negative and significant in the first specification, suggesting that vehicles have in fact become more crashworthy over time. However, once the crash test measurements are included in specifications 2 and 3, the coefficient on year designed becomes positive, small, and insignificant. Hence, there is a lack of statistical support for the assertion that vehicles have become more crashworthy apart from ways captured by the continuous measures from IIHS' moderate overlap front and side crash tests.

The results in Table 3 show that driver sorting is important. The coefficient on the IIHS badge indicator is negative and highly significant, suggesting that consumers with safer driving habits choose vehicles that are observably more crashworthy. But the impact is not very large, at least relative to the impact of the crash test measurements. In Specification 2, for example, the corresponding odds ratio on the IIHS badge indicator, 0.78, suggests that drivers of badged vehicles are about 22% less likely to die based on their driving habits alone. We observe a similar selection of safer drivers to safer vehicles when we conduct analogous exercises on an auxiliary dataset comprised of all accidents from select geographic areas, including non-fatal accidents. In particular, we find that badged vehicles are 7% less likely to be involved in an accident of any severity, a finding that is consistent with the assertion that badged vehicles are driven by more cautious drivers (see Section A.4 for details).

Next, we estimate the probability of driver death ($\hat{r}_j$) based on vehicle design. To remove the impact of driver sorting and yield fatality rates attributable to vehicle design, we set the IIHS badge indicator for every vehicle to the overall production-weighted average before predicting death rates.[29] We also remove differences in death rates attributable to vehicles aging by setting all age indicator variables to zero, and we set time fixed effects to their average. Then, we use the estimation results from Specification 3 in Table 3 to predict driver death rates ($\hat{r}_j$) attributable to vehicle design.[30]

[29] If we instead set the IIHS badge indicator to 0, then predicted fatality rates would reflect the average driving habits among consumers selecting vehicles without a safety badge. Setting the IIHS badge indicator to the production-weighted average yields predicted fatality rates for the average driver, including more cautious drivers that select badge-winning vehicles.

[30] We use the specification without vehicle weight because Anderson and Auffhammer (2013) and White (2004) have shown that aggressivity (risk of harms to occupants of other vehicles) increases with the weight

4.2 Counterfactual Ratings

In preparation for our counterfactual simulations described in Section 6.1, we rescale and mean-shift the driver death rate predictions ($\hat{r}_j$) to form a univariate continuous rating that is comparable to the binary IIHS badge rating seen by consumers in the status quo environment. Under the existing binary "badge" rating, a value of zero corresponds to the average crashworthiness among vehicles not awarded a badge, and value of one corresponds to the average crashworthiness among vehicles awarded a badge. We rescale and mean-shift the driver death rates ($\hat{r}_j$) to create a univariate continuous rating on the same scale. After this rescaling, the new continuous rating has the same scale as the status quo rating. Thus, we can simulate outcomes in counterfactual environments using the estimated demand model by replacing the binary rating with our continuous rating and finding the equilibrium.

Let $\bar{r}^{Unbadged}$ denote the production-weighted average death rate for model year 2012 to 2018 vehicles not awarded an IIHS badge: $\bar{r}^{Unbadged} = \frac{\sum_{k \notin \mathfrak{J}_{TSP}} \sum_t D_{kt}}{\sum_{k \notin \mathfrak{J}_{TSP}} \sum_t Q_{kt}} = 9.95$ quarterly fatalities per million vehicles. Similarly, let $\bar{r}^{Badged}$ denote the production-weighted average death rate for model year 2012 to 2018 vehicles that are awarded an IIHS badge: $\bar{r}^{TopSafetyPick} = \frac{\sum_{k \in \mathfrak{J}_{TSP}} \sum_t D_{kt}}{\sum_{k \in \mathfrak{J}_{TSP}} \sum_t Q_{kt}} = 7.88$ quarterly fatalities per million vehicles. We transform our driver death rate predictions ($\hat{r}_j$) to a continuous measure of crashworthiness according to the following formula: $\hat{R}_j = \frac{\hat{r}_j - \bar{r}^{Unbadged}}{\bar{r}^{Top\ Safety\ Pick} - \bar{r}^{Unbadged}}$. In Section A.4, we use a different dataset that contains all accidents in select geographic areas to estimate the risk of major injury and construct an analogous continuous measure of crashworthiness. We find the continuous measure based on risks of major injury is highly correlated ($\rho = 0.94$) with our preferred continuous measure that reflects risks of driver death.

The density of our continuous crashworthiness ratings ($\hat{R}_j$) among vehicles awarded a badge in the status quo environment are shown in Figure 5. The interquartile range (among badged vehicles) runs from 0.70 to 1.51. The corresponding interquartile range of predicted fatality rates runs from 6.82 to 8.50 quarterly fatalities per million vehicles. The highest value of the continuous safety measure is approximately 2.13, implying a predicted fatality rate of 5.54 quarterly fatalities per million vehicles, which is about half the fatality rate of the average non-badged vehicle. With better information on vehicle safety, many consumers may switch to even safer vehicles, reducing fatalities. We investigate the impact on fatalities in Section 6.

and size of the vehicle. Including weight and size in the ratings might encourage consumers to select vehicles that reduce their fatality risk but that raise fatality rates for other drivers sharing the same road network.

5 A Model of Consumer Preferences

We employ a micro-founded model of demand for automobiles, as proposed in Berry et al. (1995) and extended by Petrin (2002).[31] We assume that the conditional indirect utility individual i with demographics d derives from buying product j with characteristics ℓ in period t equals:

$$u_{ijt} = \delta_{jt} + \mu_{ijt}(x_{j\ell}, p_{jt}, \alpha_i, \nu_{i\ell}, \pi_{dit}, \theta) + \epsilon_{ijt} = \bar{u}_{it}(x_{j\ell}, p_{jt}) + \epsilon_{ijt}, \tag{3}$$

where δ_{jt} represents the mean utility for product j in period t, ϵ_{ijt} is an iid error term assumed to follow the type 1 extreme value distribution, and $\mu_{ijt}(x_{j\ell}, p_{jt}, \alpha_i, \nu_{i\ell}, \pi_{dit}, \theta)$ represents individual-specific preferences for product j given its product characteristics $x_{j\ell}$ and price p_{jt}. Specifically:

$$\mu_{ijt}(x_{j\ell}, p_{jt}, \alpha_i, \nu_{i\ell}, \pi_{dit}, \theta) = \alpha_i p_{jt} + \sum_{\ell} \sigma_l x_{j\ell} \nu_{i\ell} + \sum_{d} \sum_{\ell} \psi_{d\ell} x_{j\ell} \pi_{dit}. \tag{4}$$

$\nu_{i\ell}$ represents a standard normal draw of individual-specific preferences for characteristic ℓ, and π_{dit} denotes the value of demographic d for individual i in period t. Individual-specific price sensitivity equals:

$$\alpha_i = \frac{\alpha}{y_{it}}, \tag{5}$$

where y_{it} is consumer i's income.

The parameter set θ is comprised of σ_ℓ, $\psi_{d\ell}$, and α, which determine the extent of heterogeneity in preferences for characteristics ℓ and price.[32]

The implied market share for product j in period t equals:

$$s_{jt}(\delta_t, x, p_t, \theta) = \int \frac{\exp(\delta_{jt} + \mu_{ijt}(x_{j\ell}, p_{jt}, \alpha_i, \nu_{i\ell}, \pi_{dit}, \theta))}{1 + \sum_{k \in \mathfrak{J}} \exp(\delta_{kt} + \mu_{ikt}(x_{k\ell}, p_{kt}, \alpha_i, \nu_{i\ell}, \pi_{dit}, \theta))} f(\nu_{i\ell}, \pi_{dit}, \alpha_i) d\nu_{i\ell} d\pi_{dit} d\alpha_i, \tag{6}$$

where δ_t and p_t denote the vector of mean utilities and prices across products in period t, and x denotes a matrix of product characteristics across products.

One can invert $s_{jt}(\delta_t, x, p_t, \theta)$ to find the mean utilities (δ_t) that equate observed and predicted market shares via the contraction specified in Berry et al. (1995). The mean

[31] For an overview of current methods and norms, see Conlon and Gortmaker (2019).

[32] Assuming $\alpha_i = \frac{\alpha}{y_i}$ implies that a single parameter determines both the mean and variance of price sensitivities across consumers. Another approach is to ignore the income distribution and draw price sensitivities from a normal distribution. However, this latter approach may result in some simulated consumers having positive price sensitivities. Even a small number of drawn individuals with positive price sensitivities can strongly impact results.

utilities (δ_t) will be used to formulate some moment conditions, as described next.

5.1 Moment Condition

The moment conditions are comprised of the traditional moment conditions from random coefficient demand models and micro moments. First, one finds the component ξ_{jt} of mean utility δ_{jt} that is unexplained by a linear combination of observed product characteristics. Specifically:

$$\delta_{jt} = \sum_{\ell} x_{j\ell}\beta_{\ell} + \phi_{jt} + \eta_t + \xi_{jt}, \tag{7}$$

where ϕ_{jt} are age fixed effects (time since vehicle j's sales commenced), η_t are month fixed effects, and ξ_{jt} denotes the remaining component of mean utility that is not explained by these other factors.

The recovered error terms (ξ) and set of instruments Z are assumed to be uncorrelated, yielding a set of sample moments:

$$g_D(\theta) = \frac{1}{N}\sum_{j}\sum_{t} Z'_{jt}\xi_{jt}, \tag{8}$$

where N denotes the count of products in each market, summed across markets.

We supplement these moments with micro moments along the lines of Petrin (2002). Specifically, we include sample micro moments equaling the difference between model-predictions and empirical realizations of the covariance between demographic d_{it} and the product characteristic ℓ or purchased product j, averaging over a set of relevant periods T_m.[33] Specifically, a given micro moment m equals:

$$g_m(\theta) = \frac{1}{T_m}\sum_{t\in T_m} v_{mt} - \mathcal{V}_{mt}, \tag{9}$$

where $v_{mt} = \text{Cov}\left(d_{it}, \sum_{j\in J_t} x_{j\ell}s_{ijt}\right)$, s_{ijt} denotes the predicted probability consumer i selects product j in period t, and $\mathcal{V}_{mt}$ is the corresponding sample covariance from an auxiliary data source.

Denoting the set of micro moments g_M, the stacked set of moments is:

$$g(\theta) = \begin{bmatrix} g_D \\ g_M \end{bmatrix}. \tag{10}$$

[33] Micro moments applied to periods from January 2013 through June 2016.

The GMM objective is:

$$q(\theta) = g(\theta)'Wg(\theta), \tag{11}$$

where W is a weighting matrix.

5.2 Identification

Intuition for identification of the parameters determining the persistent variation in preferences across consumers but unexplained by demographics (σ_l, α) follows the typical arguments (Nevo, 2000). The parameters relating demographics to consumer tastes ($\psi_{d\ell}$) are identified by the added micro moments.

5.3 Instruments

Our first set of instruments are a modified version of instruments suggested in Gandhi and Houde (2019) and found to perform well in Conlon and Gortmaker (2019). Our modification accounts for the fact that vehicle availability is not binary. Rather, vehicles exit the market gradually. As vehicles are replaced by new model year variants, older model years become less attractive, have less selection on features (Copeland et al., 2011), and might not be available at local dealerships. But smaller numbers remain available for sale. Even years after a vehicle was introduced, one can still often find a new (never sold) one on a dealer's lot if willing to travel far enough. To account for this feature of the automobile market, we add weights to the competition provided by other vehicles. The weight w_{kt} for other vehicle k equals one during the first year that model k (nameplate and model year combination) is available for sale, and declines linearly to zero over the subsequent year.

Specifically, the instruments are constructed as follows, separately for each characteristic ℓ. The first instrument type equals the weighted count of other vehicles produced by the same firm with a value of characteristic ℓ within one standard deviation of vehicle j's value $\left(Z^{Same}_{j\ell t}(x) = \sum_{k\in\mathcal{J}_{ft}\setminus j} w_{kt} \times 1(|x_{j\ell} - x_{k\ell}| < \mathrm{SD}_\ell)\right)$. The second instrument type equals the corresponding weighted count of vehicles with similar values of characteristic ℓ produced by rival firms $\left(Z^{Rival}_{j\ell t}(x) = \sum_{k\notin\mathcal{J}_{ft}} w_{kt} \times 1(|x_{j\ell} - x_{k\ell}| < \mathrm{SD}_\ell)\right)$. Note that exogenous product features are fixed for a given nameplate and model year, hence variation in these instruments over time arises from changes in the set of vehicles available for sale and their age-based weights (w_{kt}). We preprocess this instrument set by first transforming the instruments using principal components, and then by keeping the smallest set of principal components accounting for 95% of the variance.

We augment the aforementioned set of instruments for markups with an instrument for cost based on exchange rates. Manufacturers are required by the American Automobile Labeling Act to report the contribution of components from country h to the final value of nameplate j, for any other country accounting for at least one third of the vehicle's value.[34] Using these data, we construct the weighted average fraction change in the exchanges rates with the United States since January 2013, weighted by the percent of the vehicle's components originating in the corresponding country: $Z_{jt}^{Exchange} = \sum_h \frac{\Psi_{jh}}{1-\Psi_{j0}} \times \frac{e-rate_t^h}{e-rate_{Jan\ 2013}^h}$, where Ψ_{jh} denotes the fraction of vehicle j's components originating from foreign country h, Ψ_{j0} denotes the share of components of unreported origin, and $e - rate_t^h$ is the exchange rate of country h's currency with U.S. dollars in period t.

The full set of instruments includes a constant, a product's exogenous characteristics $x_{j\ell}$, the modified version of the local instruments from Gandhi and Houde (2019), and the exchange rate instrument:

$$Z_{jt} = \left(1, x_{j\ell}, Z_{j\ell t}^{Same}(x), Z_{j\ell t}^{Rival}(x), Z_{jt}^{Exchange}\right). \tag{12}$$

6 Results and Counterfactuals

The estimated model coefficients shown in Table 4 are intuitive. The mean coefficient on the IIHS badge indicator is positive and significant at the 1% level. Note that although consumers can search for more detailed safety information than a badge indicator by combining ratings from multiple crash test types (each on a four-point scale), we find evidence that they focus only on the univariate rating, which is a discrete badge. See Section A.1 for details. The mean coefficients on size (W×L), miles per dollar, and indicator for recently redesigned models are also positive. The negative coefficient on horsepower divided by weight is small relative to the corresponding random coefficient, implying that a sizable share of consumers have a strong preference for vehicles with powerful acceleration.[35] The term on price is strongly significant, and the estimated mean monthly elasticity of demand is -2.57. Note that product age fixed effects for each combination of nameplate and model year are important controls to account for diminishing demand as vehicles age and the impacts of available variety (Copeland et al., 2011).

The estimated model also captures heterogeneity in preferences. Considering both random parameters and demographics-based preferences, we find heterogeneous price sensitivities and heterogeneous preferences for several product features, including crashworthiness, size, horsepower over weight, minivan indicator, and the constant. Note that heterogeneity across consumers in the constant indicates heterogeneous tastes for the inside good: new

[34] The data are available at: `https://www.nhtsa.gov/part-583-american-automobile-labeling-act-reports`.

[35] We confirm that better acceleration is generally preferred. In a representative agent model, the coefficient on horsepower divided by weight is significant when adequate controls are included. See Appendix Section A.1.

vehicles.

6.1 Counterfactuals

We begin by exploring the impact of a counterfactual ratings format—a univariate continuous rating that provides much finer information to consumers—on vehicular fatalities and welfare, holding vehicle designs fixed. Recall that we create the continuous rating by rescaling predicted driver fatality rates so that they are consistent with the binary safety ratings observed by consumers in practice and used in demand estimation. After the rescaling, a value of zero corresponds to the average driver fatality rate for unbadged vehicles (9.95 quarterly fatalities per million vehicles), and a value of 1 corresponds to the average fatality rate among vehicles awarded a safety badge (7.88 quarterly fatalities per million vehicles). See Section 4.2 for details.

To evaluate the impact of reporting precise information to consumers, we simulate purchase shares and welfare when the IIHS badge is replaced with a continuous rating. We consider both the cases where (i) prices remain fixed at levels observed in the data, and (ii) new equilibrium prices are simulated given the estimated demand system and marginal costs implied by the static multi-product firm Nash in prices equilibrium assumption.

Perhaps of greatest interest to policy makers is the impact of rating format choice on fatalities. We first simulate the log change in sales for each vehicle (nameplate and model year combination) when continuous ratings are introduced. We then plot these changes against the continuous safety rating, separately for vehicles that had received an IIHS badge and vehicles which had not (see Figure 6). Note that vehicles previously awarded an IIHS badge and revealed to be unsafe see substantial reductions in sales, as much as 50%.[36] Vehicles that were not awarded a badge but were revealed to be safer when continuous ratings are reported experience as much as a 50% increase in sales.

Next, we simulate changes in the aggregate driver fatality rate that results when consumers are presented with precise ratings and some consumers switch to safer vehicles. Specifically, we simulate the sales-weighted average fatality rate across vehicles under each ratings format, then compare. We find that a continuous rating would reduce the death rate by about 7.4% if all product features—including price—are held constant (see Table 5). For context, note that about 25,000 vehicle occupants die in crashes in the United States every year.[37] Hence, a simple back-of-the-envelope calculation suggests that reporting a continuous measure of crashworthiness would save about $0.074 \times 25,000 = 1,850$ lives per year.[38]

[36] The change in sales is primarily along the intensive margin; total vehicle sales increase by only 2.03% in the counterfactual.

[37] `https://www-fars.nhtsa.dot.gov/Main/index.aspx`

[38] The immediate impact would not be as large. The full impact would not be realized until the existing stock of vehicles (initially purchased by consumers with access only to discrete ratings) is replaced with newer vehicles purchased when continuous ratings were available.

At the value of a statistical life used by the department of transportation, \$9.4 million in 2016, this would imply a dollar benefit of about \$17.4 billion per year, despite ignoring the benefits from reducing the number and severity of injuries, which occur nearly 100 times as frequently.[39]

The simulated impact on fatalities is reduced somewhat if using simulated new equilibrium prices, rather than status quo prices. The fatality rate falls by 5.2% when discrete ratings are replaced by a continuous rating and prices adjust.[40] The smaller benefit when prices adjust arises because vehicles revealed to be substantially less safe than believed under the IIHS badge system lower their prices by about 10%, and vehicles revealed to be safer raise prices, although by a lesser amount. Specific price changes are reported in Appendix Figure A3. Note, however, that in the long run, unsafe vehicles sold at low markups may be discontinued, which may lead to a larger than 5.2% improvement in the fatality rate.

Note that a portion of the benefits of the counterfactual continuous ratings arise because the existing discrete ratings, which are constructed from a series of thresholds, can be misleading. For example a vehicle that performs exceptionally well on a number of safety dimensions but bad on just a few may not be awarded a badge, whereas another vehicle that just surpasses each of the thresholds would be awarded a badge even if it is less safe overall. Thus, existing ratings are not necessarily an ordinal representation of vehicle crashworthiness. We consider a counterfactual binary discrete rating that is based on our continuous ratings, which awards badges to the same proportion of vehicles as in the status quo environment (33.2%). Specifically, we find the continuous rating threshold for which 33.2% of vehicles yield higher continuous ratings, and assign all vehicles exceeding the threshold a safety rating equaling one. All other vehicles are awarded a safety rating of zero. We find these alternative discrete ratings would yield meaningful improvements, reducing fatalities by 3.8% (see Table 5). The benefits of an alternate discrete ratings suggests that the existing ratings were not primarily designed for their informational content, raising questions about whether the discrete rating format was chosen deliberately or whether they followed the lead of other ratings agencies in other domains. Publicly releasing the underlying data allows for competition in ratings design and format choice.

Next, we investigate the impact on consumer surplus. Some care is required. The typical multinomial logit surplus change calculation implies a comparison between ex-ante expected utility in the status quo environment where the safety characteristic is discrete with the analogous calculation where the safety characteristic is continuous. However, this is misleading because the precision of information available to consumers ex-ante differs between the two scenarios. Note that vehicle characteristics have not changed; only the

[39] See https://www.transportation.gov/resources/2015-revised-value-of-a-statistical-life-guidance and https://one.nhtsa.gov/nhtsa/whatis/planning/2020Report/2020report.html

[40] The simulated pricing equilibrium is found using the contraction proposed in Morrow and Skerlos (2011) and suggested in Conlon and Gortmaker (2019).

information about safety has. More precise ratings inform some consumers that their choice is not as crashworthy as they believed. If they still elect to purchase the same vehicle as they did in the status quo scenario, then the standard surplus calculation implies that welfare is reduced even though they are still choosing the same vehicle (at the same price), simply because they become aware of the vehicle's shortcomings. A similar problem arises for consumers who select the same vehicle in both scenarios but learn from continuous ratings that the vehicle is safer than formerly believed. In these examples, realized consumer surplus has not changed, but the standard surplus change calculation would imply that it had.

In estimating consumer surplus changes, we follow the experience goods literature. This literature differentiates between expected welfare before purchasing a product, and realized welfare after the product is consumed and true quality is observed. Thus, we can adjust the status quo welfare calculation that corresponds to ex-ante perceived welfare before purchase by subtracting changes in welfare from learning true quality after purchase.[41] Adapting the method in Reimers and Waldfogel (2019) and Train (2015), the realized consumer surplus of consumer i in period t under a specific coarse ratings equals:

$$CS_{it}^{Coarse} = \frac{\ln\left(1 + \sum_j \exp(\bar{u}_{it}(x_{j\ell}^{Coarse}, p_{jt}^{Coarse}))\right)}{|\alpha_i|} + \frac{\sum_j s_{ijt}^{Coarse}\left(\bar{u}_{it}(x_{j\ell}^{Cont}, p_{jt}^{Coarse}) - \bar{u}_{it}(x_{j\ell}^{Coarse}, p_{jt}^{Coarse})\right)}{|\alpha_i|}, \quad (13)$$

where $\bar{u}_{it}(x_{j\ell}, p_{jt})$ denotes consumer i's utility for product j in period t, excluding the idiosyncratic error term, from Equation 3, and $s_{ijt}^{Coarse} = \frac{\exp(\bar{u}_{it}(x_{j\ell}^{Coarse}, p_{jt}^{Coarse}))}{1+\sum_{k\in\mathcal{J}}\exp(\bar{u}_{it}(x_{k\ell}^{Coarse}, p_{kt}^{Coarse}))}$ denotes the probability consumer i selects product j in period t under coarse ratings. x^{Cont} denotes product characteristics under continuous safety ratings, which differs from the product characteristics under coarse ratings $\left(x^{Coarse}\right)$ only in the reported value of the safety characteristic. p^{Coarse} denotes prices under the coarse ratings. Note that the component $\frac{\sum_j s_{ijt}^{Coarse}\left(\bar{u}_{it}(x_{j\ell}^{Cont}, p_{jt}^{Coarse}) - \bar{u}_{it}(x_{j\ell}^{Coarse}, p_{jt}^{Coarse})\right)}{|\alpha_i|}$, which is not typically included in welfare calculations, accounts for changes due to learning a product's true quality after purchase. Aggregate welfare changes are calculated by integrating over the consumer types, multiplying by the market size M, and summing across periods.

The adjustment is not needed for welfare calculations in the counterfactual with more precise information; it is assumed that safety quality is observed before purchase, and therefore no new information is gleaned after purchase. The consumer surplus calculation in counterfactual environments (with precise safety information) thus follows the standard for-

[41] We acknowledge that, unlike for other experience goods or other vehicle characteristics, consumers may not fully learn how safe a vehicle is even after purchasing and driving it.

mula:

$$CS_{it}^{Cont} = \frac{\ln\left(1 + \sum_j \exp(\bar{u}_{it}(x_{j\ell}^{Cont}, p_{jt}^{Cont}))\right)}{|\alpha_i|}, \tag{14}$$

where p^{Cont} denotes prices in the counterfactual environment with precise safety information.

We find that more precise crashworthiness ratings raise consumer surplus by 7.7 billion dollars when prices remain the same in the counterfactual analyses, and 6.8 billion dollars when new equilibrium prices are simulated in the counterfactual scenario. Presumably, the costs of switching from a discrete to a univariate continuous rating are much less.[42]

The benefits of switching to a continuous rating may be muted if firms have a stronger incentive to improve safety under a discrete rating format. Although it is challenging to simulate counterfactual manufacturer investments in safety improvements for several reasons—multiple equilibria are likely, fixed costs of improving safety for vehicles sold globally cannot be inferred from U.S. market data alone, costs may change over time—we can explore whether alternative incentives can plausibly offset the gains from providing consumers with finer crashworthiness information. To that end, we (i) investigate the extent of safety reductions necessary to offset the benefits found above; (ii) explore the extent of heterogeneity in firm incentives to improve safety, which has been shown to reduce the gains from coarsening ratings (Dubey and Geanakoplos, 2010); and (iii) explore the uncertainty surrounding the optimal safety level threshold to obtain a safety badge.

To investigate the extent of manufacturer response needed to offset gains from improved information, we consider a special case where all vehicles formerly awarded a badge reduced their safety provision after switching from the status quo discrete rating to a continuous rating. Specifically, we assume that all IIHS-badged vehicles would reduce their death rate improvement over the average unbadged vehicle by x%, and we search for the value of x that would imply the same death rate results under a continuous rating. We find that IIHS-badged vehicles would need to reduce their safety advantage over the average non-badged vehicle by 23.6% to eliminate the mortality benefit arising from providing consumers with more precise safety information. Such a reaction seems large—possibly implausibly large—just from a change to the format in which ratings are presented. Furthermore, assuming all firms would reduce safety investments is presumably unrealistic. The stronger incentive of a discrete ratings' format applies only to vehicles whose safety investments would otherwise fall just below the threshold for a higher discrete score. For other vehicles, a discrete rating reduces the incentive to provide safety.

Heterogeneity across vehicles in the variable profit gained from being recognized as safe casts further doubts on the possibility that a discrete rating format could further incentivize safety provision. To formally investigate the extent of heterogeneity, we use our model to

[42] Long run changes in consumer surpluses may change if manufacturers respond to alternate ratings by redesigning vehicles recognized as safe to include other features typically valued by consumers that care about safety.

simulate the derivative of variable profits from the first 24 months of sales with respect of the level of safety reported. Specifically, $\frac{\partial \pi_j}{\partial x_{IIHS\ Badge}} = \sum_{t=\tau}^{\tau+24}(p_{jt} - c_{jt})M\frac{\partial s_{jt}(\delta_t, x, p_t, \theta)}{\partial x_{IIHS\ Badge}}$, where τ denotes the first period vehicle j is available for sale, M denotes the market size, and marginal costs c_{jt} are inferred from the multi-product firm Nash in prices equilibrium assumption.

Note the substantial variation across vehicles in the simulated impact of improving safety on variable profits that is apparent in Figure 7. The interquartile range ($17 million to $128 million) is quite large relative to the median ($48 million). Figure 7b shows a similar extent of heterogeneity among the subset of vehicles awarded an IIHS badge. Although manufacturers may elect to improve safety of some vehicles under a discrete rating, other vehicles profit less from safety recognition. A strict threshold results in some vehicles pooling with less safe vehicles, thus reducing investments in safety features. Other highly crashworthy vehicles far exceeding the threshold may reduce investments in safety provision that are not conveyed by discrete ratings and therefore go unobserved by consumers. Dubey and Geanakoplos (2010) noted that using multiple discrete ratings may somewhat alleviate this problem, but incentives to avoid pooling with less safe vehicles are less strong when there are multiple discrete rating levels.

Finally, we investigate whether uncertainty regarding the optimal threshold for a better discrete rating might undermine theoretical benefits. If the threshold is accidentally too lenient, then firms invest a small amount, just enough to pass the threshold, because any additional investments would not be observed by consumers and thus would not increase the perceived quality of their vehicles. As a result, investments in safety are reduced compared with a continuous rating. If the threshold is accidentally too strict, then firms might forgo attempting to surpass the threshold, again leading to lower investments in safety. The benefits of a discrete rating rely on properly choosing the threshold for a better discrete score.

There are two main sources of uncertainty in firms' responses. First, the rating organization does not know the exact size of the variable fixed cost investment required for a vehicle to surpass a given threshold and be recognized as safe. Second, the rating organization is uncertain of the variable profit gains manufacturers receive from being recognized as safe. Our model estimates allow us to investigate the latter: the uncertainty in variable profit gains.

We investigate the impact of uncertainty as follows. Suppose the rating organization knows the cost of increasing safety by any specific amount, but does not know the gains manufacturers receive from being recognized with a safety badge. Suppose further that the organization intends to design ratings so that vehicles whose variable profit gained is above the median will invest enough to surpass the threshold and be recognized with a safety badge. The ratings agency's goal is thus to predict the median profit gain and to choose a level of safety improvement that costs slightly less than the median profit gain. Success relies on their ability to precisely predict the median profit gain. We examine the extent of uncertainty in the median gain using the parametric bootstrapping procedure, thus assuming that our model uncertainty reflects the rating organization's perceived uncertainty. Specifically, we draw from the implied distribution of parameters. For each draw, we first calculate marginal costs implying a Nash equilibrium in prices, accounting for firm ownership Berry et al. (1995). We then calculate the derivative of profits with respect to the safety badge for each vehicle.

We find that estimated median gains to improving safety are imprecise. The interquartile range spans from \$41.9 million to \$56.7 million. The difference, \$14.8 million, is quite large relative to the median of about \$48. Thus, finding the optimal strictness of a discrete rating requires some good fortune.

Together, these mitigating factors suggest coarsening ratings may not yield substantial increases in safety investments, and may do the opposite—reduce investments in safety enhancements. Consistent with this possibility, Hunter (2020) finds that coarsened ratings reducing supplier effort in a different context. If coarsening ratings does reduce incentives for safety provision, then switching to univariate continuous crashworthiness ratings may reduce fatalities by more than 1,850.

7 Conclusion

In this paper, we investigated whether switching from the status quo discrete crashworthiness ratings to a continuous rating format would reduce fatalities and raise welfare. Our results suggest that this switch would reduce annual fatalities by 1,850 in the long run in the United States. At the current value of a statistical life used by the department of transportation, this implies that U.S. agencies should be willing to spend up to \$17.4 billion per year to make precise ratings available. We also find that precise ratings would raise consumer welfare by about \$7 billion over the five observed years. The gains are particularly large given that the costs of providing a continuous rating are presumably small. At a minimum, our results suggest that continuous ratings should be offered as an option alongside coarse ratings (currently they are not). However, the findings in Farrell et al. (2010) and Houde (2018) suggest that coarse ratings may crowd out useful information from continuous ratings, suggesting that continuous ratings should be prominently featured.

Existing theoretical work suggests that certain vehicles might have a stronger incentive to increase safety when ratings are discrete, specifically vehicles whose safety investments under a continuous ratings would be slightly below the investments needed to earn a better safety score when ratings are discrete. However, we find the requisite impact of the ratings format to be large. If all vehicles awarded a discrete safety badge reduced their safety benefit over the average unbadged vehicle by 10% after switching to a continuous rating, large mortality benefits remain. Only if *all* vehicles awarded a badge reduce safety benefits by 23% would the gains from precise information be undone. Furthermore, we find substantial heterogeneity across vehicles in the profit they gain from being recognized as safe. Thus, the optimal threshold differs substantially across vehicles. Using multiple discrete rating levels may not help, as Dubey and Geanakoplos (2010) show that incentives generated from a discrete rating decrease in the number of discrete rating levels. Finally, given substantial uncertainties as to how manufacturers will respond, the best threshold choice with available information might be far from the best threshold with perfect information, and poorly set thresholds can lead to reduced incentives to provide safety compared to continuous ratings.

We conclude that continuous ratings would likely improve welfare, suggesting that policies that enable them are preferable. Organizations can either develop such ratings themselves or allow public use of the raw data needed to develop them, as the U.S. government has with its open data website.[43] On the website, researchers are specifically encouraged to "conduct research, develop web and mobile applications, design data visualizations, and more." Opening the data to the public may be preferable to developing ratings in-house, because it allows for novel modeling innovations and other uses of the underlying data, and increases competition in the choice of how to process and present information.

Finally, our paper's implications extend beyond vehicle safety. Ratings for experience goods are abundant, and often ratings are discrete. Examples abound, from product ratings in online commerce to credit ratings (e.g., Moody's). Substantial benefits may arise from switching these ratings to a continuous format.

References

Anderson, M. and J. Magruder (2012). Learning from the crowd: Regression discontinuity estimates of the effects of an online review database. *The Economic Journal 122*(563), 957–989.

Anderson, M. L. and M. Auffhammer (2013). Pounds that kill: The external costs of vehicle weight. *Review of Economic Studies 81*(2), 535–571.

Belloni, A., V. Chernozhukov, et al. (2011). L1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics 39*(1), 82–130.

Belloni, A., V. Chernozhukov, et al. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli 19*(2), 521–547.

Berry, S., J. Levinsohn, and A. Pakes (1995). Automobile prices in market equilibrium. *Econometrica: Journal of the Econometric Society*, 841–890.

[43] See `Data.gov`.

Berry, S. T. (1994). Estimating discrete-choice models of product differentiation. *The RAND Journal of Economics 25*(2), 242–262.

Blonigen, B. A., C. R. Knittel, and A. Soderbery (2017). Keeping it fresh: Strategic product redesigns and welfare. *International Journal of Industrial Organization 53*, 170–214.

Bollinger, B., P. Leslie, and A. Sorensen (2011). Calorie posting in chain restaurants. *American Economic Journal: Economic Policy 3*(1), 91–128.

Cohen, A. and L. Einav (2003). The effects of mandatory seat belt laws on driving behavior and traffic fatalities. *Review of Economics and Statistics 85*(4), 828–843.

Conlon, C. and J. Gortmaker (2019). Best practices for differentiated products demand estimation with pyblp. *Working Paper*.

Copeland, A., W. Dunn, and G. Hall (2011). Inventories and the automobile market. *The RAND Journal of Economics 42*(1), 121–149.

Costrell, R. M. (1994). A simple model of educational standards. *The American Economic Review 84*(4), 956–971.

Dai, D. and M. Luca (2020). Digitizing disclosure: The case of restaurant hygiene scores. *American Economic Journal: Microeconomic 12*(2), 41–59.

Dranove, D. and G. Z. Jin (2010). Quality disclosure and certification: Theory and practice. *Journal of Economic Literature 48*(4), 935–963.

Dubey, P. and J. Geanakoplos (2010). Grading exams: 100, 99, 98, . . . or a, b, c? *Games and Economic Behavior 69*(1), 72–94.

Fan, Y., J. Ju, and M. Xiao (2016). Reputation premium and reputation management: Evidence from the largest e-commerce platform in china. *International Journal of Industrial Organization 46*, 63–76.

Farmer, C. M. (2005). Relationships of frontal offset crash test results to real-world driver fatality rates. *Traffic Injury Prevention 6*(1), 31–37.

Farmer, C. M. and A. K. Lund (2006). Trends over time in the risk of driver death: what if vehicle designs had not improved? *Traffic Injury Prevention 7*(4), 335–342.

Farmer, C. M. and A. K. Lund (2015). The effects of vehicle redesign on the risk of driver death. *Traffic injury prevention 16*(7), 684–690.

Farrell, J., J. K. Pappalardo, and H. Shelanski (2010). Economics at the ftc: Mergers, dominant-firm conduct, and consumer behavior. *Review of Industrial Organization 37*(4), 263–277.

Friedman, J., T. Hastie, and R. Tibshirani (2001). *The elements of statistical learning*, Volume 1. Springer Series in Statistics New York.

Gandhi, A. and J.-F. Houde (2019). Measuring substitution patterns in differentiated products industries. Technical report, National Bureau of Economic Research.

Garate, S. and P. Newbury (2019). Online reputation management through investments in quality strategically influence ratings. *Working Paper*.

Golovin, S. (2019). Technology mandates and welfare: The case of airbages. *Working Paper*.

Harless, D. W. and G. E. Hoffer (2007). Do laboratory frontal crash test programs predict driver fatality risk? evidence from within vehicle line variation in test ratings. *Accident Analysis & Prevention 39*(5), 902–913.

Hastings, J. S. and J. M. Weinstein (2008). Information, school choice, and academic achievement: Evidence from two experiments. *The Quarterly journal of economics 123*(4), 1373–1414.

Houde, S. (2018). How consumers respond to product certification and the value of energy information. *The RAND Journal of Economics 49*(2), 453–477.

Hunter, M. (2020). Chasing stars: Firms' strategic responses to online consumer ratings. *Available at SSRN 3554390*.

Jacobsen, M. R. (2013). Fuel economy and safety: The influences of vehicle class and driver behavior. *American Economic Journal: Applied Economics 5*(3), 1–26.

Jin, G. Z. and P. Leslie (2003). The effect of information on product quality: Evidence from restaurant hygiene grade cards. *The Quarterly Journal of Economics 118*(2), 409–451.

Jin, G. Z. and A. T. Sorensen (2006). Information and consumer choice: the value of publicized health plan ratings. *Journal of Health Economics 25*(2), 248–275.

Jin, Y. and S. Vasserman (2018). Buying data from consumers: The impact of monitoring programs in us auto insurance. Technical report, Working Paper.

Kullgren, A., A. Lie, and C. Tingvall (2010). Comparison between euro ncap test results and real-world crash data. *Traffic Injury Prevention 11*(6), 587–593.

Levitt, S. D. and J. Porter (2001). How dangerous are drinking drivers? *Journal of political Economy 109*(6), 1198–1237.

Lie, A. and C. Tingvall (2002). How do euro ncap results correlate with real-life injury risks? a paired comparison study of car-to-car crashes. *Traffic Injury Prevention 3*(4), 288–293.

Luca, M. (2016). Reviews, reputation, and revenue: The case of yelp. com. harvard business school nom unit working paper 12-016, 1-40.

Luca, M. and J. Smith (2013). Salience in quality disclosure: Evidence from the us news college rankings. *Journal of Economics & Management Strategy 22*(1), 58–77.

Luca, M. and J. Smith (2015). Strategic disclosure: The case of business school rankings. *Journal of Economic Behavior & Organization 112*, 17–25.

Luca, M. and G. Zervas (2016). Fake it till you make it: Reputation, competition, and yelp review fraud. *Management Science 62*(12), 3412–3427.

Mayzlin, D., Y. Dover, and J. Chevalier (2014). Promotional reviews: An empirical investigation of online review manipulation. *American Economic Review 104*(8), 2421–2455.

Metzger, K. B., S. Gruschow, D. R. Durbin, and A. E. Curry (2015). Association between ncap ratings and real-world rear seat occupant risk of injury. *Traffic Injury Prevention 16*(sup2), S146–S152.

Morrow, W. R. and S. J. Skerlos (2011). Fixed-point approaches to computing bertrand-nash equilibrium prices under mixed-logit demand. *Operations Research 59*(2), 328–345.

Nevo, A. (2000). A practitioner's guide to estimation of random-coefficients logit models of demand. *Journal of Economics & Management Strategy 9*(4), 513–548.

Peltzman, S. (1975). The effects of automobile safety regulation. *Journal of Political Economy 83*(4), 677–725.

Peterson, S., G. Hoffer, and E. Millner (1995). Are drivers of air-bag-equipped cars more aggressive? a test of the offsetting behavior hypothesis. *The Journal of Law and Economics 38*(2), 251–264.

Petrin, A. (2002). Quantifying the benefits of new products: The case of the minivan. *Journal of Political Economy 110*(4), 705–729.

Pope, D. G. (2009). Reacting to rankings: evidence from "america's best hospitals". *Journal of Health Economics 28*(6), 1154–1165.

Proserpio, D. and G. Zervas (2017). Online reputation management: Estimating the impact of management responses on consumer reviews. *Marketing Science 36*(5), 645–665.

Reimers, I. and B. R. Shiller (2019). The impacts of telematics on competition and consumer behavior in insurance. *The Journal of Law and Economics 62*(4), 613–632.

Reimers, I. and J. Waldfogel (2019). Digitization and product discovery: The causal and welfare impacts of reviews and crowd ratings.

Ryb, G. E., C. Burch, T. Kerns, P. C. Dischinger, and S. Ho (2010). Crash test ratings and real-world frontal crash outcomes: a ciren study. *Journal of Trauma and Acute Care Surgery 68*(5), 1099–1105.

Sen, A. (2001). An empirical test of the offset hypothesis. *The Journal of Law and Economics 44*(2), 481–510.

Soleymanian, M., C. B. Weinberg, and T. Zhu (2019). Sensor data and behavioral tracking: Does usage-based auto insurance benefit drivers? *Marketing Science 38*(1), 21–43.

Tadelis, S. and F. Zettelmeyer (2015). Information disclosure as a matching mechanism: Theory and evidence from a field experiment. *American Economic Review 105*(2), 886–905.

Teoh, E. R. and A. K. Lund (2011). Iihs side crash test ratings and occupant death risk in real-world crashes. *Traffic Injury Prevention 12*(5), 500–507.

Train, K. (2015). Welfare calculations in discrete choice models when anticipated and experienced attributes differ: A guide with examples. *Journal of Choice Modelling 16*, 15–22.

Traynor, T. L. (1993). The peltzman hypothesis revisited: An isolated evaluation of offsetting driver behavior. *Journal of Risk and Uncertainty 7*(2), 237–247.

U.S. Department of Transportation. Crash report sampling system. *https://www.nhtsa.gov/crash-data-systems/crash-report-sampling-system*.

U.S. Department of Transportation. Early warning reporting. *https://www.nhtsa.gov/vehicle-manufacturers/early-warning-reporting#early-warning-reporting-ewr-data-search*.

U.S. Department of Transportation. Fatality analysis reporting system. *https://www.nhtsa.gov/es/node/97996/251*.

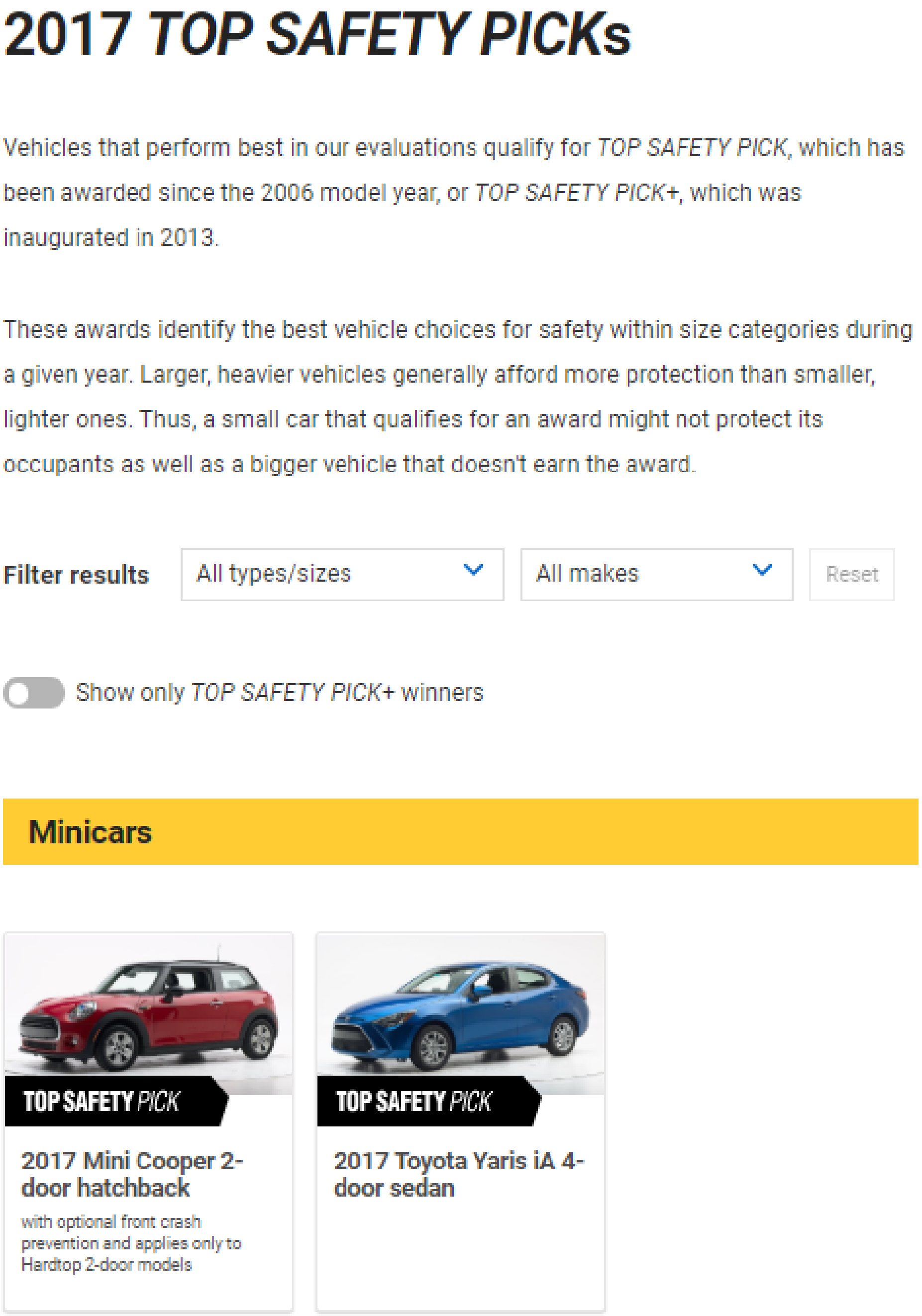
U.S. Department of Transportation (2019). Federal highway administration 2017 national household travel survey. *http://nhts.ornl.gov*.

White, M. J. (2004). The "arms race" on american roads: The effect of sport utility vehicles and pickup trucks on traffic safety. *The Journal of Law and Economics 47*(2), 333–355.

Winston, C., V. Maheshri, and F. Mannering (2006). An exploration of the offset hypothesis using disaggregate data: The case of airbags and antilock brakes. *Journal of Risk and Uncertainty 32*(2), 83–99.

Zapechelnyuk, A. (2020). Optimal quality certification. *American Economic Review: Insights 2*(2), 161–176.

Figure 1: Top Safety Picks List



2017 *TOP SAFETY PICKs*

Vehicles that perform best in our evaluations qualify for *TOP SAFETY PICK*, which has been awarded since the 2006 model year, or *TOP SAFETY PICK+*, which was inaugurated in 2013.

These awards identify the best vehicle choices for safety within size categories during a given year. Larger, heavier vehicles generally afford more protection than smaller, lighter ones. Thus, a small car that qualifies for an award might not protect its occupants as well as a bigger vehicle that doesn't earn the award.

Filter results All types/sizes All makes

Show only *TOP SAFETY PICK+* winners

Minicars

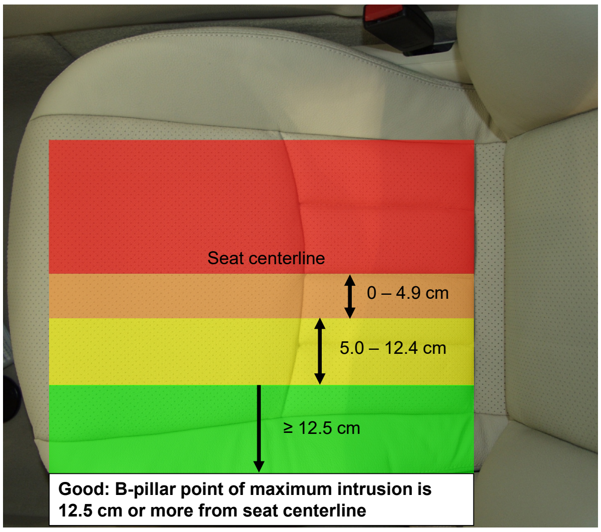
2017 Mini Cooper 2-door hatchback

with optional front crash prevention and applies only to Hardtop 2-door models

2017 Toyota Yaris iA 4-door sedan

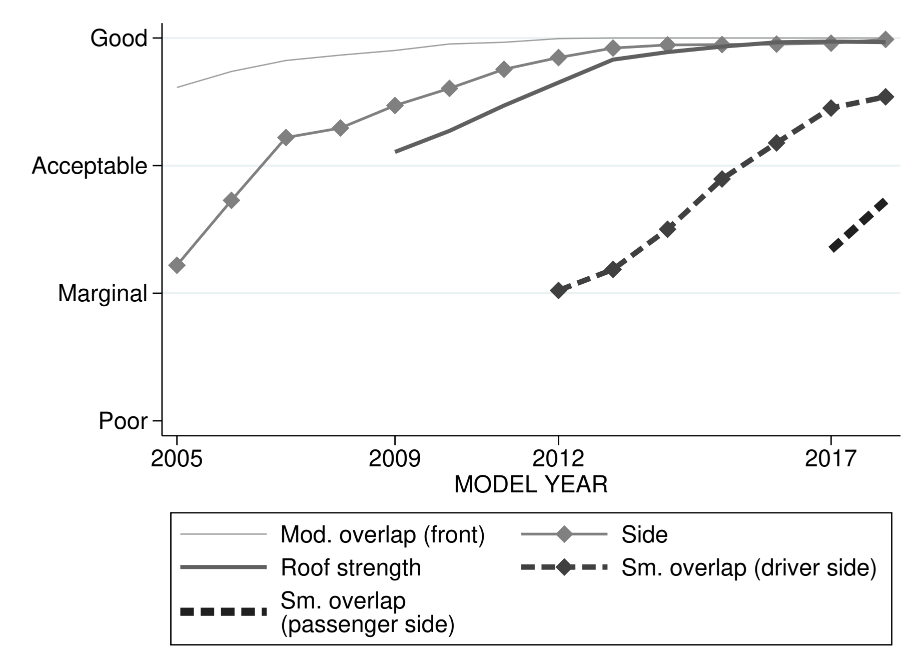
Notes: Figure shows the list of 2017 "Top Safety Picks." See: `https://www.iihs.org/ratings/top-safety-picks/2017#award-winners`. Accessed Dec 5, 2019.

Figure 2: IIHS Thresholds for B-Pillar Vehicle Structure Discrete Sub-sub-rating


Seat centerline
0 – 4.9 cm
5.0 – 12.4 cm
≥ 12.5 cm
Good: B-pillar point of maximum intrusion is 12.5 cm or more from seat centerline

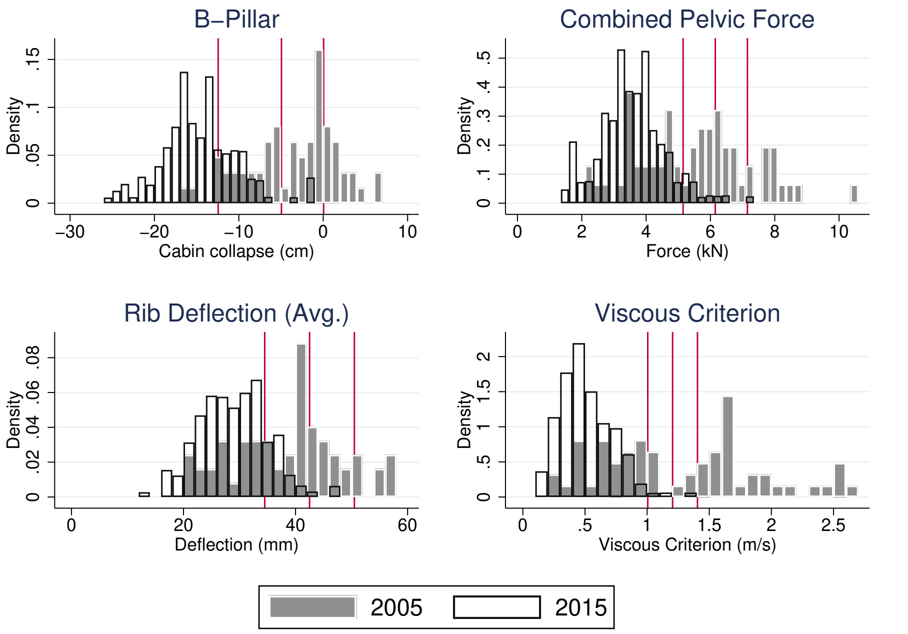
Note: The B-pillar intrusion in centimeters is translated into a discrete score based on how far the B-pillar is from the driver's seat center line following the staged crash. The driver's seat is shown from overhead in the picture. This picture was adapted from an illustration in "IIHS Side Impact Test Program – Rating Guidelines." `https://www.iihs.org/media/2104caa9-7f7e-41fa-a4a5-af65f1cab89e/l9AWAw/Ratings/Protocols/current/side_impact_guide.pdf`

Figure 3: Evolution of Discrete Subratings



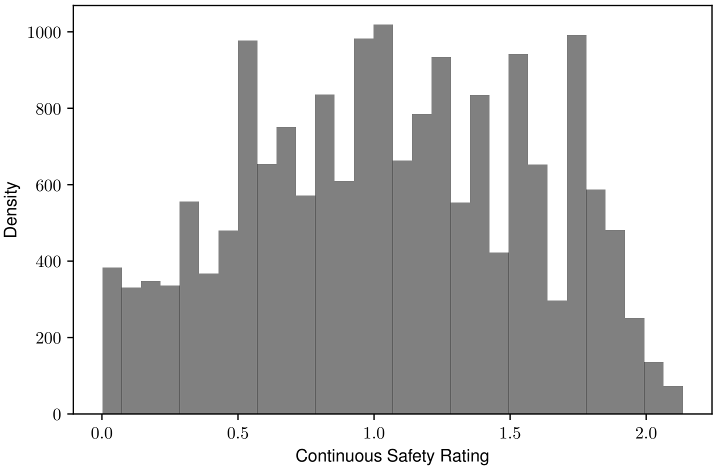
Notes: The figure shows average discrete ratings for each type of IIHS crash test for each model year between 2005 and 2018, where each is rated on a scale from 1 ("poor") to 4 ("good"). Each rating is shown from the latter of 2005 or its inception.

Figure 4: Evidence of Clumping at Thresholds



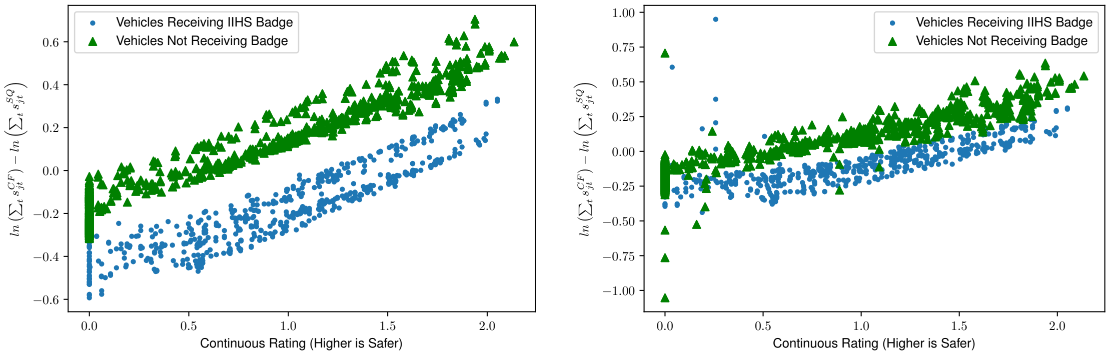
Notes: The figure shows histograms of the four discrete side impact crash test submeasures that had the lowest average discrete score on a scale from 1 (“poor”) to 4 (“good”). The discrete thresholds are depicted by vertical lines. The left-most threshold denotes the cutoff between a “good” and an “acceptable” discrete sub-sub-rating. The middle threshold denotes the cutoff between an “acceptable” and “marginal” discrete sub-sub-rating. The right-most threshold denotes the cutoff between a “marginal” and “poor” discrete sub-sub-rating

Figure 5: Distribution of Continuous Ratings Among Vehicles Earning a Safety Badge



Notes: The figure shows the density of counterfactual continuous crashworthiness ratings for vehicles earning an IIHS badge. The scale on the horizontal axis corresponds to the binary IIHS badge indicator variable. A value of 0 on the horizontal axis corresponds to the average driver fatality rate for vehicles not awarded a badge, which is 9.95 fatalities per million vehicles each quarter. A value of 1 on the horizontal axis corresponds to the average driver fatality rate for vehicles that are awarded a safety badge, which is 7.88 fatalities per million vehicles each quarter. Note that higher values along the horizontal axis correspond to lower driver fatality rates.

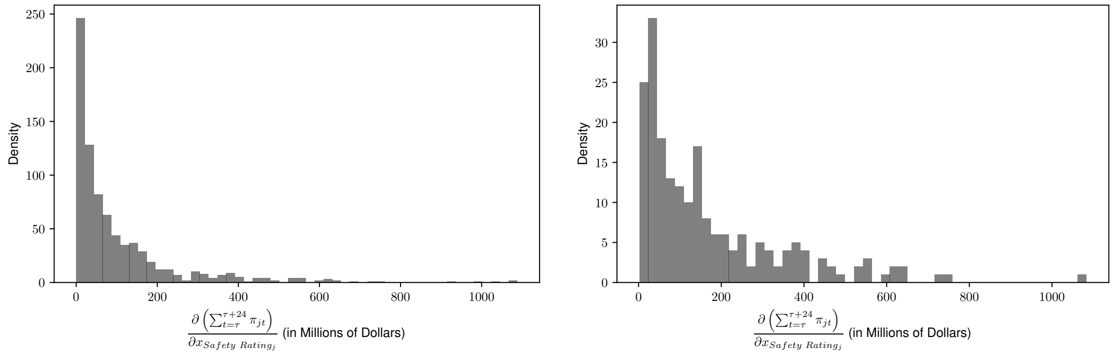
Figure 6: Change in Sales



(a) Status Quo Prices **(b)** Simulated New Equilibrium Prices

Notes: The subfigures show the log change in sales arising when a continuous safety rating is reported to consumers against the counterfactual continuous rating. Each point denotes a particular vehicle (nameplate and model year combination).

Figure 7: Impact of Safety Improvements on Profits: Distribution Across Vehicles



(a) All Vehicles **(b)** Vehicles Awarded Safety Badge

Notes: The figure shows the density of the derivative of profits with respect to the level of safety reported. Profits include all profits during the first 24 months a given vehicle (nameplate and model year) are sold. Model years 2012, 2013, 2017, and 2018, which are observed for less than 24 months, are excluded.

Table 1: Summary Statistics: Vehicle Sales and Characteristics

Model	Number		Sales[1]	1(IIHS	Size[2]			1(Major	Real
Year	Models			Badge)	(W × L)	MPD[3]	HP/WT[4]	Redesign)	Price[6]
2012	201	Avg	0.440	0.471	1.369	3.094	0.619	0.627	1.567
		SD	1.547		0.177	1.740	0.163		0.905
2013	241	Avg	3.039	0.537	1.365	3.025	0.641	0.363	1.592
		SD	6.185		0.170	1.239	0.198		0.923
2014	253	Avg	2.859	0.227	1.369	3.426	0.655	0.346	1.665
		SD	5.820		0.167	1.932	0.209		0.972
2015	258	Avg	3.050	0.290	1.367	4.217	0.667	0.299	1.673
		SD	6.043		0.163	2.090	0.213		0.948
2016	260	Avg	3.023	0.288	1.370	4.860	0.662	0.322	1.643
		SD	6.105		0.167	2.356	0.213		0.907
2017	267	Avg	3.538	0.379	1.374	4.535	0.664	0.346	1.646
		SD	6.696		0.166	2.155	0.219		0.909
2018	247	Avg	1.859	0.274	1.389	4.069	0.674	0.388	1.698
		SD	4.047		0.148	1.164	0.195		0.909

[1] Sales is quantity per month reported in thousands.

[2] Size is the product of vehicle length and width dimensions, measured in 100s of inches.

[3] MPD is miles per dollar, adjusted for inflation and reported in 1983 dollars.

[4] HP/WT is horsepower per 10 lbs of weight.

[5] 1(Major Redesign) is a binary variable indicating whether the model year is the first since a major redesign.

[6] Price is the manufacturers suggested retail price (MSRP) less any consumer incentives and rebates for elective vehicles, plus the gas guzzlers tax, where applicable. Price is adjusted for inflation and reported in tens of thousands of 1983 dollars.

Table 2: Summary Statistics: Crash Test Measurements

		Moderate Overlap Frontal Crash Test							Side Crash Test			
		Structure		Head Inj.		Neck	Chest	Leg	Chest		Leg	
Model Year		Foot Intr.	A-Pillar	HIC-15	Peak Gs	Extension	Max N_{ij}	Tibial Force	Rib Defl.	Defl. Rate	Force kN	L-M Mom.
2005	Avg	14.4	2.9	299.1	36.7	24.5	0.3	0.7	38.5	6.3	0.6	116.1
	SD	6.7	3.6	136.1	40.3	13.5	0.1	0.3	9.6	1.9	0.4	66.6
2006	Avg	12.8	2.0	295.3	35.6	22.6	0.3	0.7	36.1	5.7	0.6	113.8
	SD	5.2	2.1	124.0	39.5	10.8	0.1	0.3	9.7	2.0	0.4	65.4
2007	Avg	11.6	1.8	290.4	34.1	21.7	0.3	0.7	33.6	4.9	0.6	105.7
	SD	4.8	2.0	127.1	37.7	10.9	0.1	0.3	9.5	1.9	0.4	52.8
2008	Avg	11.2	1.7	276.4	28.9	20.6	0.3	0.7	33.1	4.8	0.6	105.2
	SD	4.6	1.9	113.4	34.2	10.4	0.1	0.3	9.2	1.8	0.3	51.4
2009	Avg	10.8	1.6	267.2	26.9	20.3	0.3	0.6	31.8	4.5	0.6	109.0
	SD	4.4	1.8	110.1	31.4	10.5	0.1	0.2	8.3	1.7	0.3	49.0
2010	Avg	10.2	1.3	260.7	24.8	19.4	0.3	0.6	31.4	4.3	0.6	112.4
	SD	4.2	1.5	111.9	31.6	10.4	0.1	0.2	7.9	1.6	0.4	51.3
2011	Avg	9.7	1.1	251.6	21.2	18.8	0.3	0.6	30.5	4.0	0.6	112.6
	SD	4.3	1.3	108.5	31.3	10.7	0.1	0.2	7.9	1.4	0.4	56.2
2012	Avg	9.3	1.0	252.7	20.0	19.0	0.3	0.6	29.5	3.9	0.6	112.7
	SD	4.1	1.0	109.5	29.7	10.9	0.1	0.2	7.3	1.4	0.4	60.1
2013	Avg	8.6	1.0	244.1	17.4	18.5	0.3	0.5	28.6	3.8	0.7	110.5
	SD	4.0	1.0	108.6	28.8	9.7	0.1	0.2	6.6	1.2	0.4	60.6
2014	Avg	8.2	0.8	237.4	15.7	18.0	0.3	0.5	28.7	3.8	0.7	102.8
	SD	4.0	0.9	104.8	26.8	9.6	0.1	0.2	6.4	1.2	0.4	58.8
2015	Avg	7.5	0.6	232.4	12.6	17.0	0.3	0.5	28.6	3.9	0.7	101.0
	SD	3.8	0.8	97.8	24.6	9.1	0.1	0.2	6.1	1.2	0.4	59.3
2016	Avg	7.1	0.6	224.4	8.5	16.9	0.3	0.5	28.9	3.9	0.7	100.1
	SD	3.8	0.8	85.3	19.9	9.2	0.1	0.2	5.8	1.2	0.3	61.8
2017	Avg	6.6	0.5	214.0	7.6	16.2	0.3	0.5	28.9	4.0	0.7	96.4
	SD	4	0.8	83.0	18.6	8.5	0.1	0.2	5.7	1.2	0.3	58.6
2018	Avg	6.4	0.5	210.9	6.5	17.1	0.3	0.5	28.3	3.9	0.7	90.0
	SD	3.6	0.7	76.7	15.9	8.3	0.1	0.1	5.1	1.2	0.3	52.5

Notes: Larger numbers denote increased injury risk and greater cabin intrusion. HIC (head injury criterion) = $\max_{t_1,t_2}\left[\frac{1}{t_2-t_1}\int_{t_1}^{t_2} a(t)dt\right]^{2.5}(t_2-t_1)$, where t_1 and t_2 denote the initial and final time (not to exceed 15 milliseconds), and a is acceleration. Gs denotes g-force. Neck extension: torque in Newton Meters (Nm). N_{ij} is a linear combination of axial loads and bending moments (see `https://www.nhtsa.gov/sites/nhtsa.dot.gov/files/rev_criteria.pdf`). Tibial axial force is measured in kilonewtons (kN). Chest deflection measures rib bone displacement, averaged across ribs. Sternum deflection denotes the rate of displacement. Femur force is measured in kilonewtons (kN). The L-M moment is measured in Newton meters (Nm).

Table 3: Relating Fatal Accident Rates to Crash-Test Measurements

	(1)	(2)	(3)	(4)	(5)
Year Redesigned	-0.0188	0.0244*	0.0157	0.0161	
	(0.0141)	(0.0130)	(0.0119)	(0.0100)	
I(Awarded Safety Badge)		-0.2500***	-0.2613***	-0.1331**	
		(0.0577)	(0.0552)	(0.0564)	
Side Crash Test Continuous Measures					
B-pillar to Driver Seat Centerline (cm)		0.0116***	0.0013	0.0102**	
		(0.0036)	(0.0031)	(0.0041)	
Neck Tension Force (kN)		-0.0192	0.1426***	0.0695	
		(0.0384)	(0.0400)	(0.0372)	
Shoulder Lateral Deflection (mm)		0.0094***	0.0031	0.0036*	
		(0.0023)	(0.0022)	(0.0021)	
Sternum Max Defl. Rate (ms)		0.0382***	0.0442***	0.0303***	
		(0.0106)	(0.0097)	(0.0095)	
Femur L-M Moment (Nm)		0.0021***	0.0013***	0.0009***	
		(0.0003)	(0.0003)	(0.0003)	
Femur A-P Moment (Nm)		0.0014**	0.0004	0.0004	
		(0.0006)	(0.0004)	(0.0003)	
Moderate Overlap Crash Test Continuous Measures					
A-pillar Rear Movement (cm)		0.0369**	0.0212*	0.0112	
		(0.0168)	(0.0127)	(0.0114)	
Head HIC-15		0.0001	0.0004**	0.0004**	
		(0.0002)	(0.0002)	(0.0002)	
Head Peak Gs		0.0009*	0.0005	-0.0002	
		(0.0005)	(0.0005)	(0.0004)	
Footwell Intr. at Footrest (cm)		0.0061	0.0161**	-0.0046	
		(0.0074)	(0.0064)	(0.0067)	
Footwell Intr. at Left (cm)		0.0137**	-0.0012	0.0097*	
		(0.0060)	(0.0061)	(0.0056)	
Footwell Intr. at Right (cm)		-0.0200***	-0.0039	0.0024	
		(0.0071)	(0.0061)	(0.0056)	
Left Femur Axial Force (kN)		-0.0320**	-0.0114	0.0031	
		(0.0153)	(0.0123)	(0.0118)	
Right Femur Axial Force (kN)		-0.0033	-0.0183	-0.0098	
		(0.0146)	(0.0118)	(0.0111)	
Right Tibia Axial Force (kN)		0.3150***	0.1965***	0.1394**	
		(0.0799)	(0.0747)	(0.0687)	
Right Foot Acceleration (g)		-0.0007	0.0006	0.0005	
		(0.0006)	(0.0005)	(0.0004)	
Vehicle Age FE	Y	Y	Y	Y	Y
Time FE	Y	Y	Y	Y	Y
Vehicle Dimensions			Y	Y	
Vehicle Type, Bodystyle, Weight				Y	
Nameplate/Design Generation FE					Y
N (in millions)	2,846	2,846	2,846	2,846	2,831
Pseudo R^2	0.0021	0.0062	0.0079	0.0095	0.0129

Notes: Reported parameters maximize the likelihood of observed driver death rates in a binary logit model. Standard errors (in parentheses) are clustered by nameplate and model-year. N = the sum of (cumulative production × quarters of exposure) across vehicle nameplates. Specification 5 excludes vehicles with zero deaths. Restricting specifications 1 to 4 to the same sample has negligible impacts on results.
* $p<0.1$, ** $p<0.05$, *** $p<0.01$

Table 4: Estimated Parameters: Random Coefficient Demand Model

Variable	Parameter Est. (Std. Err.)
Term on Price (α)	
$1/y_i$	-4.97 (0.65)***
Demographic-Based Taste Parameters ($\psi_{d\ell}$)	
Education×1(IIHS Badge)	0.21 (0.01)***
1(Rural)×1(IIHS Badge)	-0.26 (0.01)***
1(Rural)×Size (W×L)	2.76 (0.07)***
Has Child×Size (W×L)	1.51 (0.07)***
Has Child×1(Minivan)	2.36 (0.11)***
Standard Deviations (σ_ℓ)	
Constant	10.30 (1.07)***
HP/WT	1.57 (0.54)***
Mean Coefficients (β_ℓ)	
1(IIHS Badge)	0.26 (0.07)***
Size (W×L)	0.33 (0.45)
MPD	3.57 (2.24)
HP/WT	-0.48 (0.58)
1(Recently Redesigned)	0.07 (0.06)
Vehicle Age FE	Y
Vehicle Type FE	Y
Month FE	Y

Notes: MPD denotes miles per dollar, and HP/WT denotes horsepower divided by weight. Vehicle age denotes the number of months since sales of the vehicle (nameplate, model-year) commenced. Vehicle type indicates style (e.g., SUV, sedan), and an indicator for being classified as a luxury or executive vehicle. The model was estimated using two-step GMM. Standard errors are clustered by nameplate and model year combinations.
* p<0.1, ** p<0.05, *** p<0.01

Table 5: Simulated Impacts of a Alternate Rating Formats

	Continuous Ratings		CF Discrete
	Status Quo Prices	Simulated New Equilibrium Prices	
%Δ in	-7.399%	-5.207%	-3.786%
Driver Deaths	[-10.110, -4.787]	[-8.050, -2.535]	[-5.919, -1.795]
Δ in	\$7.669	\$6.819	\$5.818
Consumer	[4.281, 14.611]	[3.867, 13.178]	[3.315, 11.096]
Surplus			
(in Billions)			

Notes: Reported changes are relative to the status quo discrete rating. The death rate is defined as the death rate per vehicle (deaths/total purchases). The 95% confidence interval is computed via parametric bootstrapping and is shown in brackets.

A Online Appendix

A.1 Representative Agent Model

In this section, we investigate which vehicle features and reported safety information components impact consumer choices using a representative agent logit model. We use the Berry (1994) linearization, and we instrument for price using the modified Gandhi and Houde (2019) instruments detailed in Section 5.3 (See Appendix Table A1). Most parameter estimates are intuitively sensible. The coefficient on IIHS badge is positive and significant, confirming that on average consumers have a preference for safer vehicles. Most other parameters have their anticipated sign or are insignificant.

The IIHS prominently features "Top Safety Pick" badges, suggesting that this may be the most conspicuous rating to consumers. But there are finer ratings available on the IIHS's website. Specifically, on some of its site's subpages, the IIHS reports a separate rating for each crash test type (moderate overlap front, small overlap front, side, and roof strength) on a four-point scale (poor [1]; marginal [2]; acceptable [3]; good [4]). To investigate whether consumers use information from these specific tests when making vehicle purchase decisions, we include the average rating across the specific crash tests as an additional explanatory variable in the last column of Appendix Table A1. Note that the coefficient on this variable is small and statically insignificant, whereas the coefficient on the "Top Safety Pick" indicator remains large and highly significant, confirming that consumers focus predominantly on the "Top Safety Pick" badges that are prominently shown on the IIHS's website (see, for example, Figure 1).

Table A1: Estimated Parameters: Representative Agent Demand Model

	Dependent Variable is $log(s_{jt}) - log(s_{0t})$				
	(i)	(ii)	(iii)	(iv)	(v)
1(IIHS Badge)	0.448***	0.448***	0.450***	0.319***	0.305***
	(0.107)	(0.106)	(0.106)	(0.0696)	(0.0722)
1(Major Redesign)	0.163***	0.141**	0.140**	0.132**	0.132**
	(0.059)	(0.056)	(0.055)	(0.051)	(0.051)
Price (α)	-1.122***	-1.031***	-1.024***	-2.880*	-2.904*
	(0.181)	(0.186)	(0.185)	(1.543)	(1.550)
Size (W × L)	2.973***	2.546***	2.553***	4.302**	4.293**
	(0.392)	(0.481)	(0.481)	(2.107)	(2.111)
MPD	-2.337	0.777	1.848	-7.651***	-7.492***
	(3.733)	(3.568)	(3.885)	(2.554)	(2.641)
HP/WT	-0.232	-0.294	-0.297	1.902**	1.910**
	(0.575)	(0.585)	(0.584)	(0.822)	(0.824)
1(4WL or AWL)	-0.186	-0.281*	-0.284*	-0.202	-0.202
	(0.164)	(0.159)	(0.159)	(0.278)	(0.279)
Avg. of Specific Tests					0.0460
(Scale from 1 to 4)					(0.0604)
Model Year FE	Y	Y	Y	Y	Y
Vehicle Age FE	Y	Y	Y	Y	Y
Engine Type FE	Y	Y	Y	Y	Y
Vehicle Type FE		Y	Y		
Month FE			Y	Y	Y
Nameplate FE				Y	Y

Notes: Table A1 shows estimation results under various specifications of a representative agent logit model. Estimates were obtained by regressing $log(s_{jt}) - log(s_{0t})$ on product characteristics and instrumenting for price via two-stage least squares regression (Berry, 1994). MPD denotes miles per dollar. Vehicle age denotes the number of months since sales of the vehicle (nameplate, model-year) commenced. Engine type is either gas, electric, or hybrid. Vehicle type indicates style (e.g., SUV, sedan). Avg. of Specific Tests denotes the average across specific tests (moderate overlap front, small overlap front, side, and roof strength), each of which is rated 1 ("poor"), 2 ("marginal"), 3 ("acceptable"), or 4 ("good"). Standard errors, clustered by nameplate and model year combinations, are shown in parentheses.
* $p<0.1$, ** $p<0.05$, *** $p<0.01$

A.2 Supplementary Information

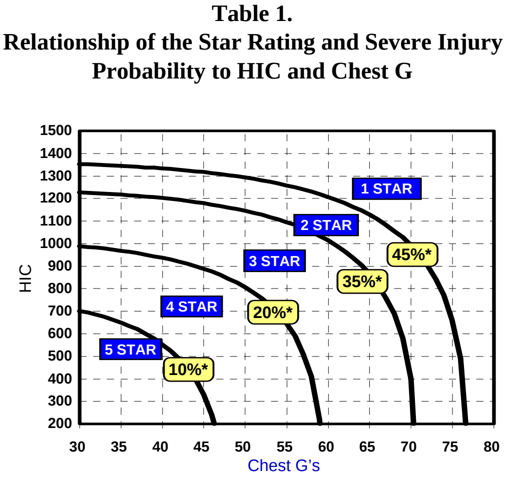


Table 1.
Relationship of the Star Rating and Severe Injury Probability to HIC and Chest G

Figure A1: NHTSA: Discrete Ratings Thresholds

Notes: The figure shows the mapping between the two injury ratings and the discrete frontal-collision crashworthiness rating from the NHTSA, prior to 2011 (afterwards additional injury measures were included, precluding two-dimensional depictions). The HIC denotes the head injury criterion, and the Chest G's denotes the g-force applied to the dummy's chest in the staged collision. The percentages denote overall risk of serious injury. Source: "New Car Assessment Program (NCAP): Past, Present, and Future," `https://www-nrd.nhtsa.dot.gov/pdf/esv/esv17/Proceed/00245.pdf`.

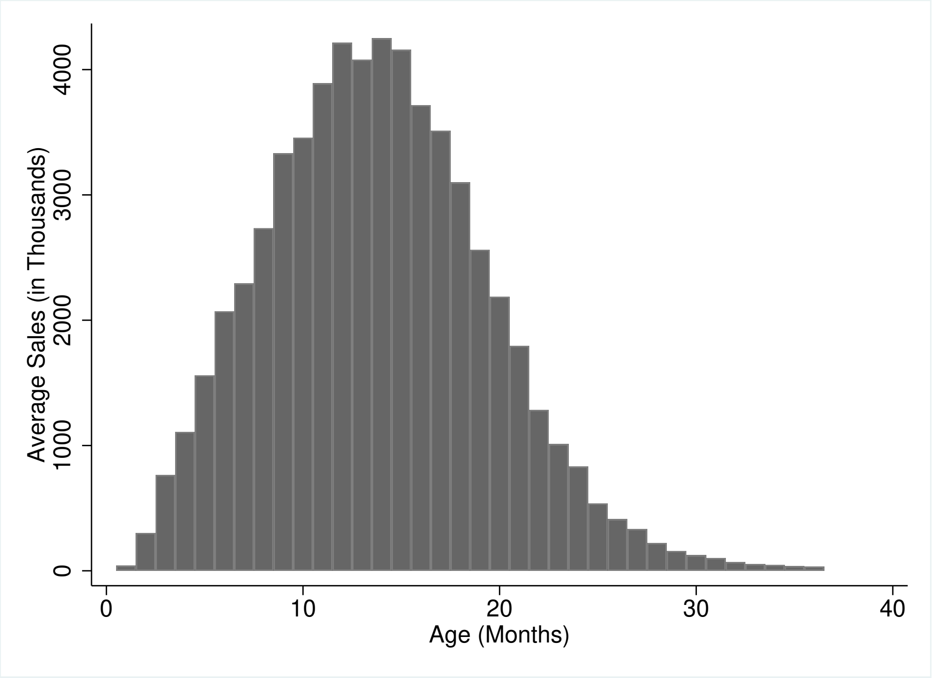


Figure A2: Average Sales vs. Age

Notes: The figure shows the evolution of sales from the time the vehicle is introduced in the market. The scale on the horizontal axis corresponds to age in months since introduction. The vertical axis corresponds to average sales volume. Vehicle sales beyond 36 months are excluded.

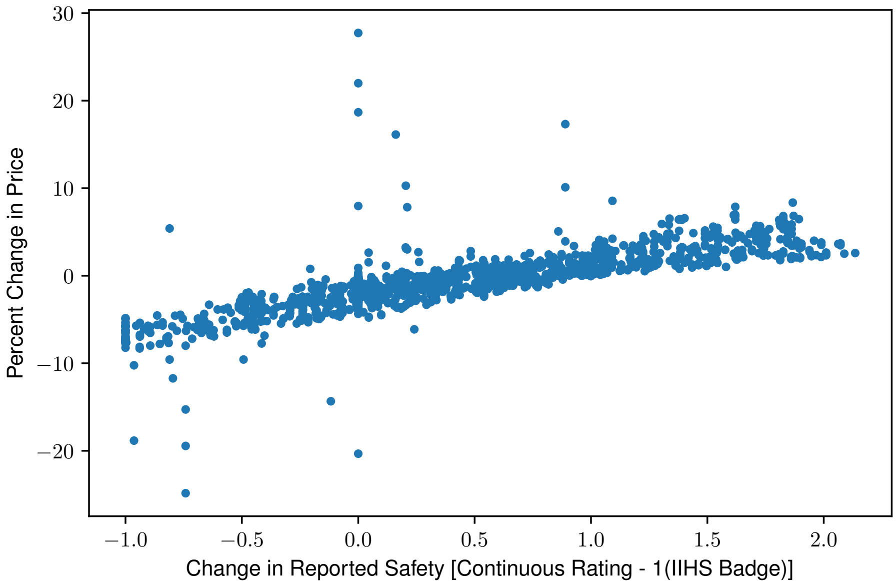


Figure A3: Changes in Equilibrium Prices When Continuous Ratings Are Reported

Notes: The figure shows the predicted change in average price (across observed periods) against the change in reported safety: continuous rating - 1(IIHS Badge). Each point represents a single vehicle (nameplate and model year combination).

Table A2: Mapping of Subratings to Side-Impact Crashworthiness Rating [IIHS]

	Assigned Demerits if Subrating Equals:			
	Good	Acceptable	Marginal	Poor
Vehicle Structure	0	2	6	10
Driver				
Head Protection	0	2	4	10
Head and Neck	0	2	10	20
Torso	0	2	10	20
Pelvis and Femur	0	2	6	10
Rear Passenger				
Head Protection	0	2	4	10
Head and Neck	0	2	10	20
Torso	0	2	10	20
Pelvis and Femur	0	2	6	10
Overall Side Rating Demerit Ranges	0-6	8-20	22-32	34+

Notes: This table is an adapted version of the table in IIHS's "Side Impact Crashworthiness Evaluation—Weighting Principles for Vehicle Ratings." The final discrete score for the IIHS side crash test reported to consumers depends on the total demerits accumulated based on the discrete subratings. Subratings themselves have sub-sub-ratings. For example, the torso subrating depends on five sub-sub-ratings (peak rib deflection, average deflection [across ribs], viscous criterion [ribs], rib deflection rate, and shoulder deflection), each of which is assigned a discrete score (good, acceptable, marginal, or poor) based on whether the measure exceeds certain thresholds. Each subrating equals the worst of its sub-sub-ratings. If less than 6 demerits are assigned across the various subratings, the overall discrete side crash rating is "good." Between 8 and 20 demerits corresponds to an "acceptable" side crash rating. And so on.

A.3 Two-Vehicle Crashes Fatality Analyses

Earlier papers have taken a different approach to addressing the main omitted variable problem. Specifically, drivers with safer (or less safe) driving habits may disproportionately choose vehicles with better observed discrete crashworthiness ratings. If not addressed, one might attribute the impacts of safer driving habits to the inherent crashworthiness of rated vehicles.

The most promising previous approach to this problem exploited Newton's 3rd law of motion using two-vehicle accidents: in an accident, both vehicles exert equal force on each other. Suppose that the fatality risk for a driver in an accident follows the binary logit formula: $R_j = \exp(\alpha+\beta X_j)/[1+\exp(\alpha+\beta X_j)]$, where X_j are variables potentially impacting crashworthiness of vehicle j. Then, following Farmer (2005), the conditional probability that the driver of vehicle 1 died given that exactly one of the drivers died equals:

$$R_1 = \exp(\beta(X_1 - X_2))/[1 + \exp(\beta(X_1 - X_2))].$$

Note that the unit of observation in these analyses is a particular vehicle in an accident. The number of observations equals twice the number of two vehicle accidents in which exactly one driver died. In our sample of the FARS data, we had 3,741 two vehicle accidents in which exactly one driver died, and for which we have information on the crashworthiness measures and vehicle characteristics for both vehicles in the accident.

We augment this strategy in two ways. First, we include many vehicle features as explanatory variables, including continuous measures from staged crash tests. Second, we include the difference in the IIHS badge indicator. Recall that after controlling for the continuous measures from staged crash tests that collectively determine whether a vehicle is awarded an IIHS badge, the IIHS badge indicator captures the marginal impact of driver selection. Suppose that the strategy proposed used by Farmer (2005) fully accounts for selection of drivers with safer driving habits. Then the IIHS badge indicator that accounts for remaining selection after controlling for the many continuous measures from the staged crash tests should not explain which driver dies. If instead it remains significant, it suggests that driver selection remains an important confounder. This may result if the more aggressively driven vehicle is in a more precarious situation after deflecting from the initial collision. For example, maybe the driver of a speeding vehicle does not die from the initial collision but instead succumbs afterwards if remaining speed causes the vehicle to roll over or careen off the road, implying that the two vehicles in the same accident were not subject to the same conditions.

The results are show in Table A3. Note that the coefficient on the IIHS badge remains negative and highly significant, suggesting that selection of drivers with safer habits may confound the two-vehicle analyses of vehicle crashworthiness in prominent papers in the literature (Farmer, 2005; Kullgren et al., 2010; Lie and Tingvall, 2002).

Table A3: Fatality Analyses Using Two-Vehicle Accidents

	Coefficient (Std. Err.)
Difference Between Vehicles in:	
Year Redesigned	0.018 (0.019)
I(Awarded Safety Badge)	−0.207 (0.058) $***$
Frontal Crash Test Measures	
Difference Between Vehicles in:	
A-pillar Rear Movement (cm)	0.048 (0.017) $***$
Head HIC-15	0.000 (0.000)
Head Peak Gs	0.005 (0.001) $***$
Neck Bending Moment	0.001 (0.003)
Neck Max Force N_{ij}	−0.823 (0.304) $***$
Footwell Intr. at Center (cm)	−0.009 (0.007)
Left Foot Acceleration (Gs)	−0.003 (0.001) $***$
Right Femur Axial Force (kN)	0.004 (0.020)
Right Tibia Axial Force (kN)	0.537 (0.127) $***$
Side Crash Test Measures	
Difference Between Vehicles in:	
B-pillar Distance from Centerline (cm)	0.019 (0.005) $***$
Neck Tension Force (kN)	0.058 (0.067)
Sternum Max Defl. Rate (ms)	0.028 (0.020)
Sternum Avg. Deflection (mm)	−0.004 (0.004)
Femur L-M Moment (Nm)	−0.001 (0.000)
Femur A-P Moment (Nm)	−0.000 (0.001)
Left Femur Force (kN)	0.307 (0.073) $***$
Vehicle Characteristics	
Difference Between Vehicles in:	
Length (in)	0.006 (0.003) $**$
Width (in)	0.020 (0.018)
Height (in)	−0.059 (0.007) $***$
Weight (100 lbs)	−0.043 (0.010) $***$
Horse Power	−0.002 (0.001) $**$
Engine Size (liters)	−0.227 (0.082) $***$
Vehicle Age	0.012 (0.004) $***$
Observations	7481

Notes: Standard errors shown in the right column, in parentheses.
* $p<0.1$, ** $p<0.05$, *** $p<0.01$

A.4 Supplemental Crash Test Analyses Using CRSS Data

In this section, we supplement the analysis of vehicle crashworthiness in Section 4.1 by considering additional measures of adverse events in crashes. In Section 4.1, we focused on driver fatalities using the FARS dataset, which excludes non-fatal accidents. In this section, we use a different dataset to explore the relationship between vehicle characteristics and other adverse outcomes, such as accident frequency, likelihood of minor injury, and likelihood of major injury. We could in principle construct alternative continuous measures to the ones described in Section 4.1 that instead reflect the vehicle's ability to protect occupants from major injury. Additionally, we can examine whether the estimated coefficient on IIHS badge indicator—purportedly a measure of driver selection—is consistent with driver selection as the mechanism when we focus on the propensity of a vehicle to be in an accident of any severity.

For this analysis, we utilize CRSS data, which include multiple measures of driver injuries—minor, major or fatal—for all vehicular crashes occurring within 60 sampling sites within the United States between 2011 and 2017. Based on these data, we can construct the likelihood of an adverse outcome in a collision:

$$P(y_{jt} = 1) = \frac{exp(x_{j\ell t}\gamma_{\ell t})}{1 + exp(x_{j\ell t}\gamma_{\ell t})}. \tag{15}$$

That is, conditional on a crash, we can model each measure of driver injury y_{jt} as a logistic regression. Similarly, we can also model the unconditional propensity of a vehicular crash similar to equation 2:

$$LL = \sum_j \sum_t D_{jt} \times ln\left(F\left(x_{j\ell t}|\gamma_{\ell t}\right)\right) + \left(Q_{jt} - D_{jt}\right) \times ln\left(1 - F\left(x_{j\ell t}|\gamma_{\ell t}\right)\right), \tag{16}$$

where D_{jt} is the number of vehicular crashes of nameplate-year j in period t and Q_{jt} is the cumulative production of j at t.[44]

The results are shown in Table A4. Column 1 indicates the unconditional propensity of being in a vehicular accident, while columns 2–4 show the likelihood of injury conditional on an accident. We regress these on the same set of regressors as Table 3 to maintain consistency. Most of the continuous crash measurements follow the same sign as the main fatality analysis.

We note that driver selection, measured by the coefficient on the safety badge dummy, persists to affect other outcomes of crashworthiness as well. Vehicles awarded a safety badge are estimated to be 7% less likely to be involved into a crash. Conditional on being in an

[44] Because the CRSS samples only 60 sites across the United States, we scale overall cumulative production numbers by the yearly proportion of driver fatalities from these sites (on average 3% of driver fatalities are attributed to these sites) to create the representative cumulative production numbers for these areas.

accident, badged vehicles are 5.6% less likely to sustain a major injury and 17.5% less likely to die. These results—in particular the first—are consistent with safer driving practices being a confounder for the "true" (mechanical) crashworthiness, indicating it needs to be controlled for.

Next, in Table A5 we also create alternate measures of crashworthiness based on the specifications from Table A4 above. We find that most measures bear a high correlation with our preferred measure of crashworthiness based on unconditional driver fatalities, from Section 4.2.

Table A4: Relating Fatal Accident Rates to Crash Test Measurements

	Unconditional	Conditional on Crash		
	Propensity of Crash	Any Injury	At Least Major Injury	Death
Year Redesigned	0.0309***	0.0018	-0.0030	0.0010
	(0.0060)	(0.0046)	(0.0093)	(0.0258)
I(Awarded Safety Badge)	-0.0721**	0.0096	-0.0578*	-0.1796*
	(0.0285)	(0.0176)	(0.0333)	(0.1053)
B-pillar to Driver Seat	0.0070***	0.0039**	0.0032	-0.0033
Centerline (cm)	(0.0022)	(0.0019)	(0.0035)	(0.0102)
Neck Tension Force (kN)	0.0266	-0.0331*	0.0194	0.2065**
	(0.0236)	(0.0193)	(0.0353)	(0.1045)
Sternum Max Defl. Rate (ms)	0.0250***	0.0298***	0.0458***	-0.0024
	(0.0083)	(0.0056)	(0.0099)	(0.0327)
Shoulder Lateral Deflection (mm)	0.0012	-0.0008	-0.0034*	-0.0050
	(0.0017)	(0.0010)	(0.0018)	(0.0052)
Femur L-M Moment (Nm)	0.0006***	-0.0001	0.0005**	0.0012*
	(0.0002)	(0.0001)	(0.0002)	(0.0007)
Femur A-P Moment (Nm)	0.0002	-0.0001	-0.0001	-0.0005
	(0.0002)	(0.0002)	(0.0003)	(0.0007)
A-pillar Rear Movement (cm)	-0.0113	0.0023	0.0269**	0.0556*
	(0.0076)	(0.0075)	(0.0125)	(0.0332)
Head HIC-15	-0.0002	0.0001	0.0002	-0.0003
	(0.0001)	(0.0001)	(0.0002)	(0.0005)
Head Peak Gs	-0.0004	0.0000	0.0006	0.0016
	(0.0003)	(0.0003)	(0.0005)	(0.0013)
Footwell Intr. at Footrest (cm)	0.0046	-0.0087**	-0.0001	0.0125
	(0.0052)	(0.0037)	(0.0070)	(0.0214)
Footwell Intr. at Left (cm)	0.0001	0.0091***	0.0014	0.0030
	(0.0044)	(0.0032)	(0.0062)	(0.0166)
Footwell Intr. at Right (cm)	-0.0084**	0.0048*	-0.0021	-0.0142
	(0.0040)	(0.0028)	(0.0051)	(0.0146)
Left Femur Axial Force (kN)	-0.0117*	0.0099	-0.0044	-0.0086
	(0.0066)	(0.0070)	(0.0128)	(0.0324)
Right Femur Axial Force (kN)	-0.0207**	-0.0035	0.0184	0.0346
	(0.0089)	(0.0060)	(0.0118)	(0.0350)
Right Tibia Axial Force (kN)	-0.0182	0.0625	0.1826**	0.2146
	(0.0595)	(0.0401)	(0.0802)	(0.2340)
Right Foot Acceleration (g)	0.0003	-0.0006**	-0.0004	-0.0017
	(0.0003)	(0.0003)	(0.0005)	(0.0015)
Vehicle Age FE	Y	Y	Y	Y
Time FE	Y	Y	Y	Y
Observations	58,759,565	142,527	142,527	142,527

Standard errors are shown in parentheses.
* p<0.1, ** p<0.05, *** p<0.01

Table A5: Correlating Predicted Crashworthiness From Crash Outcomes and Driver Fatalities

	CRSS Outcomes			
		Conditional on Crash		
	Propensity of Accident	Any Injury	At least Major Injury	Fatality
FARS Fatality	-0.0308	-0.1429	0.9451	0.9111