Are Coarse Ratings Fine? Examining the Impacts of Format Choice for Crashworthiness Ratings*

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

[†]Brandeis University

[‡]Monash University

— PRELIMINARY —

Jan 25, 2020

Abstract

Many ratings organizations intentionally coarsen ratings before public presentation, for example using a discrete badge rather than a continuous rating. We investigate the impact of coarsening ratings empirically in the context of automobile crashworthiness ratings. Specifically, we estimate a random coefficient model of demand for safety under status quo coarse ratings, then simulate the outcomes under a counterfactual scenario with a novel continuous crashworthiness rating system. We find the finer crashworthiness ratings would alter consumer vehicle purchases and thereby reduce fatalities by about 10%, implying about 2,550 fewer fatalities in the United States annually. We then explore whether reduced incentives to provide crashworthy vehicles under a continuous rating scheme might offset these benefits. We find that the requisite manufacturer response is implausibly large. We conclude that a continuous rating scheme would increase consumer welfare and save lives.

*We would like to thank Aiden Zhang for superb research assistance. We also would like Gary Feld and Leo Yon for generously helping us acquire needed data for this project. Lastly, we would like to thank import.io for providing a discounted rate for their web-scraping software.

1 Introduction

In addition to gathering, verifying, and providing relevant information to consumers, certification and rating organizations can choose how to package and present the information. Many of them intentionally coarsen information before public presentation. For example, the two U.S. organizations that evaluate crashworthiness (safety) of vehicle nameplates employ a small number of discrete ratings. A natural question is whether intentionally coarsening the ratings further promotes their mission to reduce deaths, injuries, and economic losses from traffic accidents. In the context of crashworthiness ratings, we investigate whether a continuous rating would reduce fatalities and increase consumer welfare, relative to the 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 than coarse ratings.[1] But, we argue that firms 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, at least in theory.

Even if coarse ratings are optimal in theory, certifiers may lack the information or expertise needed to design ratings to optimally encourage quality improvements. In the context of crashworthiness ratings, safety certifiers appear to follow the medical literature, but not the economics literature, when designing coarse rating thresholds. Even with the relevant expertise, certifiers may be unable to precisely predict how manufacturers will respond to the ratings thresholds due to inherent uncertainties. The risk of poorly setting thresholds for coarse rating may outweigh any theoretical gains from coarsening ratings.

The economics literature has examined mechanisms policy makers can use to yield desired outcomes. For example, government organizations can require certain features (Golovin, 2019). A less heavy-handed approach is to report quality ratings to consumers. Empirical papers in the literature have typically found that that ratings do influence choices (Anderson and Magruder, 2012; Hastings and Weinstein, 2008; Jin and Leslie, 2003; Luca, 2016; Tadelis and Zettelmeyer, 2015), and that firms attempt to strategically manipulate quality ratings (Fan et al., 2016; Garate and Newbury, 2019; Luca and Zervas, 2016; Mayzlin et al., 2014; Proserpio and Zervas, 2017). For an overview of the quality certification literature, see Dranove and Jin (2010). However, the literature on the impacts of the ratings format choices is relatively sparse. In this paper, we study whether using continuous quality rating improves outcomes.

Vehicle crashworthiness ratings are an auspicious context for studying this question.

[1] While bounded rationality might limit comprehension of continuous ratings, the results in Farrell et al. (2010) and Houde (2018) suggests consumers do meaningfully respond to continuous ratings, and that continuous ratings appear more informative even after accounting for mental processing limitations.

The gains from improved safety ratings may be large: vehicle accidents take a significant economic toll of $242 billion (in 2010) and are the 11^{th} leading cause of death in the US.[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 counterfactual rating schemes, we first construct continuous crashworthiness ratings by relating fatalities to vehicle features and measures obtained from staged crash tests, yielding vehicle specific probabilities of driver death. Next, we estimate demand for vehicles features (including safety) under status quo coarse ratings using a random coefficient discrete choice model of consumer demand. We then simulate consumer choices and resulting death rates under a continuous crashworthiness rating, and evaluate whether it is plausible that a continuous rating would reduce incentives for safety provision by enough to offset the gains from providing finer information to consumers.

When constructing our continuous measure of vehicle crashworthiness, we address a potential selection issue typically ignored in the literature, e.g. (Farmer, 2005) — drivers that care more about safety might both drive more carefully and choose vehicles with better crashworthiness ratings, conflating the impacts of safety features and driving behaviors.[3,4] To address these selection issues we include both continuous measures of crashworthiness (which are not observed by consumers) as well as the observable discrete crashworthiness measures as explanatory variables. Intuitively, this approach is similar to a regression discontinuity design. 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 cars is meaningful. By estimating the discontinuity, we can remove it (and the impact of driver selection) from our continuous crashworthiness ratings. We also employ a novel approach to control for several other confounders typically ignored in the literature: age/deterioration, trends in road safety conditions, and unrelated safety improvements.

Our estimates show that continuous ratings provide a great deal more information, leading consumers to substantially alter vehicle purchase decisions in counterfactual simulations. Our point estimates imply that switching from the status quo discrete ratings to a continuous rating would reduce fatal accidents by about 10%, implying an annual reduction of 2,550 fatalities in the United States. Due to the imprecision of our estimates, however, the

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

[3] The opposite selection issue could arise as well, as risky drivers who are more likely to be involved in a serious accident might benefit more from safety features and therefore choose cars with better crashworthiness ratings. Observed safety features and ratings may also directly impact chosen driving behaviors (Cohen and Einav, 2003; Peltzman, 1975).

[4] Some recent studies attempt to limit endogeneity issues by focusing on the relative risks of injuries between occupants of each vehicle involved in two-vehicle accidents (Kullgren et al., 2010) However, this strategy might underestimate the crashworthiness of higher-rated vehicles if safety features such as well-designed crumple zones reduce decelerations thereby limiting injury risk of occupants in both cars, including the vehicle collided with.

ninety five percent confidence interval ranges from a 0% change in fatalities to a 39% reduction in fatalities. It is possible that manufacturers would respond to continuous ratings by lowering investments in crashworthiness. However, based on our point estimates, we find that safer rated vehicles would need to collectively reduce the their safety improvement over vehicles not recognized as safe by 32% in order to offset the reduction in fatalities arising from improved consumer sorting with finer ratings. We view this large of a response to be implausible.

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 construction of a 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.

2 Background

2.1 Ratings Organizations

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

A major and growing focus of both organizations is crash test safety ratings.[5] Starting in the 1990s, the sole NHTSA and IIHS crash test was a frontal collision into a fixed barrier.[6] The NHTSA and IIHS added side barrier tests in 1996, and 2003, respectively.[7] Based on concerns related to rollover deaths, in 2000, the NHTSA added a measure of the probability of rollover, and in 2009, the IIHS added a roof strength test. In 2012, the IIHS added a more stringent frontal crash test where only 25% of the front on the driver side collides with the barrier, and in 2017 they expanded this test to the passenger side. Both organizations note that comparisons of frontal collision test are only valid between cars 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 are typically based on optional packages or select trims.

Ratings are highly visible. The ratings from the IIHS and NHTSA are prominently shown

[5]Timelines: A timeline of NHTSA's Safety Ratings Program (`https://www.nhtsa.gov/ratings`) , About the Institute, Milestones (`http://www.iihs.org/iihs/about-us/milestones`).

[6]The NHTSA uses a fully head-on collision, the IIHS uses a moderate overlap collision where only 40% of the car's front hit the barrier.

[7]The NHTSA side-crash test involves a car-like object crashing into the vehicle, the IIHS involves a SUV/truck like vehicle that hits the side higher up.

in car reviews on websites like ConsumerReports.com, Edmunds.com, and USNews.com.

2.2 Format of Ratings

Consumers appear to focus on discrete measures of crashworthiness reported by the NHTSA and IIHS. The many continuous measurements from crash tests are only reported in inconspicuous technical reports, and understanding the many continuous scores would be challenging and time-consuming for consumers. So both agencies transform the continuous measurements into simplified scores that are featured on their respective websites and publications. The NHTSA reports separate discrete scores for each crash test type (e.g. side, front) on five star scale. The IIHS reports a discrete four point scale for each test type, 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 Figures 1 and 2), and "Top Safety Pick" badges are prominently shown on pages for specific vehicles (see Figure 3).

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

The vehicle structure sub-rating in the side-impact crash test is based on a single continuous measurement, the final location of the cars 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 instrusion is mapped to a discrete sub-rating according to the criteria in Figure 4.

The injury sub-ratings for each body region (head, neck, torso, pelvis/femur) are intuitively similar. Each body region sub-rating equals the worst sub-rating implied by the various continuous measurements for that region.[9]

Finally, the five sub-ratings (vehicle structure, and injury regions: head, neck, torso, pelvis/femur) are combined into an overall side-impact crash test rating using a demerit based system depicted in Table 1.[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.

[8] "IIHS Side Impact Test Program – Rating Guidelines." `https://tinyurl.com/y7rg4azf`

[9] "Side Impact Crashworthiness Evaluation – Guidelines for Rating Injury Measures." `https://tinyurl.com/yak22xm9`

[10] "Side Impact Crashworthiness Evaluation – Weighting Principles for Vehicle Ratings." `https://tinyurl.com/y9kmh6z5`

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

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. 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 test.[12]

2.3 Anecdotal Evidence of Manufacturer Response

The response of manufacturers to new crash test categories provides strong suggestive evidence that manufacturers design cars with crashworthiness ratings in mind. Figure 5 shows the evolution of average IIHS crashworthiness rating (on a four point scale from 1 ["poor"] to 4 ["good"]), separately by crash type and model-year. Note that when new tests are introduced average ratings are initially rather low, but improve quickly, suggesting manufacturers respond. Note, for example, that in 2012, when the driver-side small overlap test was introduced, the average score was about two (marginal). In response, manufacturers redesigned the structure of 97 different models for the US market, mostly by using stronger or thicker materials for the cabin.[13] By 2016, three fourths subsequently scored "good," the best possible rating.[14] However, the IIHS remained suspicious that manufacturers were intentionally focusing on safety improvements captured by the driver-side crash test, and were ignoring the untested passenger side. After confirming these suspicions using a small sample of vehicles, the IIHS added the passenger-side small overlap test to their protocols in 2017. As Figure 1 shows, average scores on the passenger 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.[15]

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 introduced (in 2003) just prior to the beginning of our sample, and uses a smaller number (16) of underlying continuous 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 which had the lowest average assigned discrete scores in 2005, thus having the most room for improvement. Figure 6 shows histograms of these continuous sub-measures, in 2005, and a decade later. Note that

[12] `https://tinyurl.com/y8dv8vcq`

[13] https://tinyurl.com/yc63qu7d

[14] ibid.

[15] Among cars with a passenger ride rating in 2017, the average passenger side ratings on a four point scale (from 1=poor to 4=good) was 2.3, much lower than the corresponding driver side rating (3.4).

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 sub-ratings. In 2015, the clumping is more pronounced near the lowest threshold, which corresponds to the highest ratings. We view this clumping as strong evidence that manufacturers respond to tests, especially considering that there are factors which should mitigate such clumping: (i) there are many sub-measures included in the rating which are simultaneously impacted by vehicle design choices, and (ii) the IIHS is not the only crash test organization (although is arguably the most stringent), as there are other domestic (NHTSA) and international (e.g. Europe's NCAP and Japan's JNCAP) ratings agencies which may also influence manufacturer's design choices.

2.4 Ratings Format and Welfare

We use the Hotelling model framework to provide intuitive arguments 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. Imagine two firms located at the endpoints of a linear city. The distance/ travel cost between a given consumer's location along the linear city and a firm's location represents horizontal product differentiation along the constellation of non-safety features. We assume that that safety is a vertical product feature – all consumers agree that safer vehicles are better. However, we allow for heterogeneous tastes for safety, e.g., some consumers value safety more than others. Specifically, we are interested in the case where consumers near the midpoint of the linear city have higher or lower valuations for safety than consumers near the endpoints of the linear city, who are captive consumers of one firm or the other.

When choosing safety levels of their vehicles, firms will trade off the costs of improving crashworthiness with the corresponding increase in revenues. The extent to which revenues increase as safety improves depends the value for safety among consumers who are nearly indifferent between the two firms. Firms will invest more in safety if marginal consumers have higher valuations for safety. And vice versa. How much captive consumers (located near the ends of the linear city) value safety may be irrelevant — Small changes in crashworthiness may not impact their product choice.

However, a utilitarian welfare maximizer will trade off the costs of improving safety with the total benefit to all consumers, including those who are captive consumers of one of the firms. This suggests that if marginal consumers have much higher valuations for safety than captive consumers then firms may provide more than the welfare maximizing level of safety investments in their vehicles. By contrast, if marginal consumers have relatively low valuations for safety, firms may provide less than the welfare maximizing levels of investments in safety features. Perfect information conveyed through continuous crashworthiness ratings

does not guarantee that firms will provide welfare maximzing levels of safety features.

Now suppose a safety rating organization reports only a binary measure indicating whether some arbitrary safety threshold has been surpassed. Suppose the marginal consumer places more value on safety than the average consumer, and firms exert more than the welfare maximizing level of effort into improving safety when a continuous safety measure is reported. A regulator could improve welfare by reporting a binary measure of whether the welfare maximizing level of safety has been exceeded. Firms would not be recognized for exerting effort beyond the threshold, and thus would have no incentive to do so.

Perhaps a more interesting question is whether a binary safety threshold could induce firms to invest more in safety compared to when a continuous measure is reported. Firms whose effort under continuous ratings would be near but still below the threshold after coarsening have a choice. If they increase their effort slightly, they will be recognized as high quality after ratings are coarsened. Otherwise, after coarsening they will be pooled with unsafe vehicles and recognized as unsafe. Costrell (1994) shows such agents below but near the threshold increase effort when ratings of their effort are coarsened. However, agents lack incentives to exert effort beyond the coarse ratings threshold, since their additional effort would not be observed. Therefore, some agents who already exceeded the threshold may reduce their effort after ratings are coarsened. One might improve welfare further by adding additional discrete ratings for such agents to strive for. But using too many thresholds for different categories 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. The first concern is that thresholds for discrete ratings may be poorly set, even if it is possible to choose the ideal threshold. 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, firms will under invest in safety features. If the threshold is set too high, firms may forgo attempts to surpass it. A related concern arises if there is inherent uncertainty as to how firms will respond. In that case, the risks of poorly setting a threshold may outweigh the small gains achieved if the right threshold is chosen. Expected safety improvements may be lower under even an ex-ante optimal coarse ratings scheme.

3 Data

We construct two different datasets for separate analyses. The first set of data are used to estimate demand for vehicles. They contain static vehicle characteristics (from WARDS Auto), monthly real prices [including time-varying consumer incentives (Auto News Database) and rebates/gas guzzler taxes] measured in tens of thousands of 1983 dollars, monthly sales in the United states by nameplate and model year (provided by the Alliance of Automobile

Manufacturers), and an indicator for 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 observations of a vehicle to the first 24 months of sales, and censor monthly sales below 25. The dataset spans the period ranging from January 2013 to December 2017 (60 months). Additionally, the demand analysis uses information on the real income distribution from the U.S. census, again measured in tens of thousands of 1983 dollars.[17] We assume the market size in a given period equals the number of households in the United States divided by the 12 months in a year.

Trends in vehicle characteristics are summarized in Table 2. Note that the fraction of vehicles awarded an IIHS badge fluctuated substantially, in part due to evolving standards.[18] However, a typical consumer may have been unaware of these evolving criteria, as the changes were not prominently noted.

The second dataset which is used to estimate continuous crashworthiness ratings includes information on the IIHS safety badge status and crash test measurements for 2057 distinct vehicles (nameplate and model-year combinations) between model years 2005 and 2018, from the IIHS website `https://www.iihs.org`. The measurements are comprised of a large set of recorded extent of cabin intrusion and dummy injury severity. Trends in select crash test measurements are reported in Table 3. Note that the measures have declined in magnitude over time, implying lower levels of injury and less cabin intrusion, at least on average. There were many other measurements from crash tests considered in our analyses as well, although the others were found to have minimal impacts on driver deaths. See Section 4 for details.

These crashworthiness measurements are combined with with characteristics (e.g. dimensions, weight, etc.) from WARDS Auto, quarterly production data from the Early Warning Reporting database (see the TREADS Act), and quarterly driver fatalities from the Fatality Analysis Reporting System (FARS). The resulting dataset contains time-invariant vehicle characteristics and crash measurements along with quarterly fatalities and cumulative production, separately for each combination of nameplate and model year.[19] The fatality dataset spans a period quarters from 2005 through 2017.

[16] We exclude passenger vans and vehicles with manufacturers suggested retail prices exceeding $50,000 measured in 1983 dollars (about $130,000 in current dollars).

[17] https://www.census.gov/data/tables/time-series/demo/income-poverty/historical-income-households.html

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

[19] 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.

4 A Model of Vehicle Crashworthiness

4.1 Constructing Continuous Crashworthiness Ratings

In order to evaluate counterfactual scenarios with a univariate continuous crashworthiness rating — which did not exist — we must first construct such a rating. In this section, we construct a continuous crashworthiness rating by relating driver deaths in vehicle j occurring in period t to static vehicle characteristics and crash test measurements and time-varying controls in a binary logit model.[20] Let the probability a driver of vehicle j dies equal $F(x_{j\ell t}, \gamma_l) = \frac{\exp(\sum_\ell x_{j\ell t}\gamma_l)}{1+\exp(\sum x_{j\ell t}\gamma_l)}$, where $x_{j\ell t}$ denotes vehicle characteristics, crash test measurements, safety ratings, and other controls, and γ_l denotes parameters to be estimated. The log-likelihood function is:

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

Where D_{jt} denotes the number 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 confounders which may be correlated with crash test measurements, but may independently influence driver fatality rates. If wear and tear reduces crashworthiness, we would observe higher fatality rates for older cars.[21] 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. Similarly, if road safety improves over time, then vehicles released later face safer road conditions on average. Because vehicles released later also have tend to have better crash test measurements, we risk conflating the impact of improved vehicle designs and road conditions. 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 estimated of the influence crash test measurements on fatality rates we must control for the year each vehicle was designed.

A feature of the auto industry can be exploited to control for the aforementioned confounders; age, changing road conditions, and other vehicle safety improvements. Blonigen et al. (2017) notes that changes between model-years are typically superficial. Major redesigns occur only about every 5 years. In fact, the IIHS exploits this feature, assigning

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

[21] We remain agnostic to the reason vehicles have more driver deaths as they age. In unreported analyses using public traffic citations data, we found 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.

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, 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 that point, since 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 — e.g. a consumer that is unusually concerned about safety might both choose to drive unusually cautiously and buy a vehicle with good observed crashworthiness ratings. If not addressed, one might conflate the impacts of driver's habits and crash tests measurements on driver death rates.

Our identification strategy exploits the discrete nature reported ratings. Suppose, for example, that two nearly identical vehicles earn identical ratings on all crash test measures except one. Suppose one of the vehicle's cabin intrusion in the B-Pillar test is measured to be 12.4 cm, the other measures 12.5 cm. While these measurements are very similar, implying nearly identical crashworthiness, the measurements are on opposite sides of the rating threshold, implying only one of the vehicles may be awarded a "Top Safety Pick" (i.e., recognized as safe). This might lead drivers who care more about safety, and who might be inherently safer (or less safe) drivers, 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 safer drivers. Hence, differences in driver death rates between vehicles with the same observed 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, 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 measurments of injury and cabin intrusion from staged crash tests. Because the former, the discrete score, reflects the influence of

sorting of consumers with safer driving habits, we must remove its impact before using the model to predict a vehicle's inherent ability to protect its occupants. Specifically, we include the IIHS badge status as an explanatory variable in our model. Then, when predicting a vehicles' inherent crashworthiness using the model, we reassign the value of the IIHS badge to the average value across vehicles, thereby removing impacts of driver composition on predicted fatality rates.

In Table 4, 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 4 includes only age and time fixed effects and the year the vehicle was last redesigned. Next, we consider the 26 continuous measurements of cabin instrusion and dummy injury from the moderate-overlap frontal crash test, and the 16 continuous measurements from the side-overlap test, as well as the discrete safety score (has safety badge/does not). To alleviate concerns of overfitting, we incorporate and estimate a LASSO penalty parameter using 10-fold cross validation. We only apply the penalty to the continuous crash-test measurements, and restricting their coefficients to have the anticipated (weakly positive) sign, implying more instrusion and more force to the crash test dummy raises fatality risk.[22] 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).[23] Thus, we report the unpenalized binary logistic model with the eleven selected variables in Column 2. In the third column, we add vehicle weight and dimensions as additional explanatory variables. In Column 4, we include fixed effects for each combination of make, nameplate, and design period (e.g. all model years between major resigns). These fixed effects approximately capture all differences in crashworthiness across vehicles.

Next we compare the relative values of the pseudo R-squared across models. Adding safety variables increase the pseudo R-squared from 0.0023 to 0.0063. Adding in size and dimensions increases the pseudo R-squared to 0.0084. 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 0.0125, 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 are attributable to continuous crash test measurements from the front and side crash tests and basic vehicle characteristics.

[22] We exclude continuous measurements from crash tests introduced later.

[23] Also see discussion in Section 3.8.5 in Friedman et al. (2001), https://tinyurl.com/ydxzc4um

Note that the coefficient on year designed is 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 smaller and insignificant. Hence, there is a lack of statistical support for the assertion that vehicles have become more crashworthy in ways that are not captured by the continuous measures from IIHS' moderate overlap front and side crash tests.

The results in Table 4 show driver sorting is important. The coefficient on IIHS badge indicator is negative and highly significant, suggesting 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. Take, for example, Specification 2. The corresponding odds ratio on the IIHS badge indicator, 0.83, suggests that drivers of badged vehicles are about 17% less likely die based on their driving habits alone.

To investigate the impact of crash test measurements on fatalities, we simulate the impact of a one standard deviation worsening along each continuous crash test safety measurement, using specification 2 in Table 4. We predict such worsening would increase driver fatalities in this sample of vehicles by 88%.

Next, we estimate the probability of driver death ($\hat{r}_j$) based on vehicle design. To remove the impact of driver sorting, we set the IIHS badge indicator for every vehicle to the overall average (fraction of vehicles earning the IIHS badge) before using our model to predict death rates. We also remove differences in death rates attributable to vehicles aging by setting all age indicator variables to zero. Then, we use the estimation results from Specification 2 in Table 4 — except with time fixed effects set to their average — to predict driver death rates ($\hat{r}_j$) attributable to vehicle design.[24]

Before simulating counterfactual demand with continuous safety ratings in Section 6.1, we must first rescale predicted death rates so they are consistent with the binary measure of safety used in demand estimation (1 if receiving a "Top Safety Pick" or "Top Safety Pick+" badge, and 0 otherwise). I.e. we need to rescale death rates ($\hat{r}_j$) so that a value of 0 corresponds to the production-weighted average crashworthiness of vehicles not awarded an IIHS badge and a value of 1 corresponds to the production-weighted average crashworthiness of vehicles awarded an IIHS badge.

Let $\bar{r}^{Unbadged}$ denote the sales-weighted average death rate for vehicles not awarded an IIHS badge: $\bar{r}^{Unbadged} = \frac{\sum_{k \notin \mathcal{J}_{TSP}} \sum_t D_{kt}}{\sum_{k \notin \mathcal{J}_{TSP}} \sum_t Q_{kt}} = 11.9$ fatalities per million vehicles each quarter. Similarly, let $\bar{r}^{Badged}$ denote the sales-weighted average death rate for vehicles that are awarded an IIHS badge: $\bar{r}^{TopSafetyPick} = \frac{\sum_{k \in \mathcal{J}_{TSP}} \sum_t D_{kt}}{\sum_{k \in TSP} \sum_t Q_{kt}} = 7.6$ fatalities per million vehicles each quarter.

[24] We use the specification without vehicle characteristics because Anderson and Auffhammer (2013) has shown that aggressivity (risk of harms to occupants of other vehicles) increases with the weight of the vehicle. Including weight in the ratings might encourage consumers to select vehicles that reduce their fatality risk, but raise fatality rates for other drivers sharing the same road network.

Our rescaled continuous measure of crashworthiness equals: $\hat{r}_j^{rescaled} = \frac{\hat{r}_j - \bar{r}^{Unbadged}}{\bar{r}^{Top\ Safety\ Pick} - \bar{r}^{Unbadged}}$. In counterfactual simulations, the IIHS badge indicator variable is replaced with the rescaled continuous measure of safety ($\hat{r}_j^{rescaled}$) for vehicles that were awarded an IIHS badge, censoring values less than zero.

The density of our continuous crashworthiness ratings among vehicles awarded a safety badge — vehicles whose ratings will change under a counterfactual with a continuous measure of crashworthiness — are shown in Figure 7. The average continuous rating is about 0.81, the interquartile range is 0.56 to 1.14, and the highest value of the continuous safety measure is approximately 1.53. Note that some vehicles that earned a safety badge are barely safer, if at all, than the average vehicle not awarded a badge. With better information on vehicle safety, many consumers may switch to even safer vehicles, reducing fatalities. We investigate the extent to which this occurs in the next two sections.

4.2 Inferring Mechanisms for Improving Safety

A related question relates to the mechanisms by which manufacturers improve crashworthiness of their vehicles. Are safety improvements yielded by adding more to the vehicle (e.g., more airbags, more steel in safety cage), implying that improving crashworthiness raises variable costs? Or are safety improvements at least partially driven by vehicle design (e.g., crumple zones), implying safety raises fixed costs? 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 then regress an indicator for newly acquiring a safety badge on and indicator denoting whether the model year was a redesign, yielding the following results:

$$\begin{array}{rcccccc} 1(Newly\ Earned\ Badge_{jt}) & = & 0.32 & + & 0.11 \times I(Newly\ Redesigned_{jt}) & + & \epsilon_{jt} \\ & & (0.01) & & (0.01) & & \end{array} \quad (2)$$

The results of this linear probability model imply that nameplates which 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 14% of the time, an 11 percentage point increase.[25] Hence, much of the observed safety improvements coincided with major redesigns, suggesting 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.

[25]Similar results were obtained in a analogous regression that included model year fixed effects to account for variation in badge requirement stringency.

5 A Model of Consumer Preferences

5.1 Demand

We employ a micro-founded model of demand for automobiles, as proposed in Berry et al. (1995). We assume the conditional indirect utility individual i derives from buying product j with characteristics ℓ in the current period t equals:

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

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 $\mu_{ijt}(x_{j\ell}, p_{jt}, \nu_{i\ell}, \alpha_i, \theta)$ represents individual-specific preferences for product j given its product characteristics $x_{j\ell}$ and price p_{jt}. Specifically:

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

where $\nu_{i\ell}$ are individual-specific preferences for each characteristic ℓ drawn from the standard normal distribution, and individual-specific price sensitivity equals:

$$\alpha_i = \frac{\alpha}{y_i} \tag{5}$$

where y_i is consumer i's income, drawn from the distribution of income from the United States census.

The parameter set θ is comprised of σ_ℓ and α, which determine the extent of heterogeneity in preferences for characteristics ℓ and price, respectively.[26]

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

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

where δ_t and p_t denote the vector of mean utilities and prices across products in period t.

One can invert $s_{jt}(\delta_t, x, p_t, \theta)$ to find the mean utilities (δ_t) which equate observed and

[26] Assuming $\alpha_i = \frac{\alpha}{y_i}$ implies that a single parameter determines both the mean and variance of price sensitivities. 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. In unreported exercises using the latter approach, we found that less than 5 percent of simulated consumers had positive price sensitivities at estimated parameter values, but that these consumers were drastically over-represented among simulated consumers choosing to purchase a vehicle.

predicted market shares, via the contraction specified in Berry et al. (1995). The mean utilities (δ_t) will be used to formulate the moment conditions, as described in Section 5.2

5.2 Moment Condition

The moment conditions are formed from the mean utilities (δ) obtained from the demand model. The errors ξ_{jt} represent the component unexplained by a linear combination of observed product characteristics:

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

where x denotes the set of observed characteristics influencing utility, ϕ_{jt} denotes time since vehicle j's sales commenced (i.e., age), and ξ_{jt} denotes the remaining mean utility 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:

$$g(\theta) = Z'\xi \tag{8}$$

The corresponding GMM objective is:

$$q(\theta) = g(\theta)'Wg(\theta) \tag{9}$$

where W is a weighting matrix.

5.3 Instruments

We employ the instruments suggested in Gandhi and Houde (2019).[27] The instruments are constructed as follows, separately for each characteristic ℓ. The first instrument generated using characteristic ℓ equals the count of other vehicles produced by the same firm with a value of characteristic ℓ within one standard deviation of vehicle j's value $\left(Z_{j\ell t}^{Same}(x) = \sum_{k \in \mathcal{J}_{ft} \setminus j} 1(|d_{jk\ell}| < \text{SD}_\ell)\right)$. The second instrument generated using characteristic ℓ equals the corresponding count of vehicles with similar values of characteristic ℓ produced by rival firms $\left(Z_{j\ell t}^{Rival}(x) = \sum_{k \notin \mathcal{J}_{ft}} 1(|d_{jk\ell}| < \text{SD}_\ell)\right)$. 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. The full set of instruments for the demand-side model includes a constant, a product's exogenous characteristics

[27] Conlon and Gortmaker (2019) notes that these instruments perform better than classical instruments.

$x_{j\ell}$, and the local instruments from Gandhi and Houde (2019):

$$Z_{jt} = \left(1, x_{j\ell}, Z_{j\ell t}^{Same}(x), Z_{j\ell t}^{Rival}(x)\right) \tag{10}$$

We pre-process the instrument set (Z) by first transforming the instruments using principal components, and then keeping the smallest set of principal components accounting for 99% of the variance. Finally, we rescale each by dividing by the Euclidian norm.

6 Results and Counterfactuals

Before presenting results from the full random coefficient model of consumer demand, we begin with estimates from various specifications of a representative agent logit model in Table 5. Product age fixed effects for each combination of nameplate and model year are included in all specifications to account for diminishing demand as vehicles age and the impacts of available variety on dealers' lots (Copeland et al., 2011). 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 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 sites 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 for demand in the last column of Table 5. Note that the coefficient on the 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 conspicuous "Top Safety Pick" badges that are prominently shown on the IIHS's website (see Figures 1, 2, and 3).

The results from the random coefficient model are presented in Table 6. Note the heterogeneous-taste parameters for horsepower and four wheel drive (includes all wheel drive) are statistically significant, indicating consumers have heterogeneous tastes for these product features. Likewise, the scaling parameter on the random coefficient for IIHS badge indicator is large and statistically significant, indicating that consumer have heterogeneous tastes for vehicle crashworthiness. Recall from Section 2.4 that heterogeneous tastes for safety imply that firms may over- or under-provide safety when a continuous rating system is used. In later counterfactual analyses, we will explore the impacts of the coarsening

observed in practice.

Interpreting the mean coefficients for variables that also have a random coefficient takes care. In particular, note that the negative mean coefficient on safety badge in the heterogeneous agent model is compatible with the positive coefficient on safety badge in the representative agent model. In the random coefficient model, a vehicle's purchasers are predominately comprised of simulated consumers that have strong preferences for that vehicle's prominent characteristics (Berry et al., 1999). Vehicles with large values of horsepower over weight sell disproportionately to consumers with a strong preference for acceleration. Likewise, vehicles with IIHS badges sell predominately to consumers that value safety. Even though the mean coefficient on safety is negative, the large random coefficient implies that there are still many consumers that highly value safety, especially considering the fact that market shares of individual vehicles are typically low (the outside good share is 0.877 on average). Earning a badge yields strong demand from these consumers, resulting in higher overall sales. The negative mean coefficient on safety and large random coefficient merely rationalizes the strong cross-price elasticities between vehicles recognized as safe and market shares that are only somewhat higher than vehicles not recognized as safe.

6.1 Counterfactuals

In this subsection, we explore the impact of a different ratings format on welfare and vehicular fatalities. Specifically, we simulate the impacts of a switch from the discrete ratings used in practice to a counterfactual scenario where a continuous rating of crashworthiness is reported to consumers, assuming vehicle design and prices remain fixed. We then explore whether it is plausible that manufacturers would respond to a continuous rating scheme by reducing crashworthiness enough to offset these gains.

Perhaps of greatest interest to policy makers is the impact of ratings' format choice on fatalities. We simulate the aggregate driver death rate under both status quo discrete ratings and a counterfactual scenario with a continuous crashworthiness rating by summing the product of predicted shares $\hat{s}_{jt}$ and predicted driver death rates $\hat{r}_j$, and dividing by the predicted share electing for the inside good. Specifically, for each scenario, the simulated aggregate driver death rate equals: $\frac{\sum_t \sum_j (\hat{r}_j \times \hat{s}_{jt})}{\sum_t \sum_j \hat{s}_{jt}}$.[28]

We find that a continuous rating would reduce the death rate by 10.2%. See Table 7. For context, note that about 25,000 vehicle occupants die in crashes in the United States every year.[29] Hence, a simple back of the envelope calculation suggests that reporting a continuous measure of crashworthiness would save about $0.102 \times 25,000 = 2,550$ lives per

[28] We focus on death rate per vehicle rather than total deaths because the latter is influenced by the share of consumers choosing the inside good. It is difficult to interpret an increase in new vehicle purchases on overall deaths, because we do not observe which outside option consumers would opt for if they do not purchase a new vehicle (e.g., keep older and less safe vehicle) nor the fatality rates for outside options.

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

year.[30]

We also consider the impact of ratings format on consumer surplus. Consumer surplus is assumed to equal the average consumer surplus across individuals and observed time periods: $CS = \frac{1}{NT}\sum_t \sum_i \left(\ln\left(1+\sum_j \exp(\bar{u}_{ijt})\right)/\alpha_i\right)$. The simulated percent change in consumer surplus in the automobile market is 28.9%. See Table 7.

The benefits of switching to a continuous rating may be muted if firms have a stronger incentive to improve safety under a discrete ratings format. While it is challenging to simulate counterfactual manufacturer investments in safety improvements for several reasons — multiple equilibria are possible, inferring variable costs assuming Nash in prices equilibria may be inappropriate with fleet-wide fuel regulations, fixed costs of improving safety for vehicles sold globally cannot be inferred from US market data alone — we can explore whether alternative incentives can plausibly offset the gains from providing consumers with finer crashworthiness information. To that end, we consider a special case where all vehicles with positive continuous ratings decrease crashworthiness by the same proportion when a continuous rating is used, and resimulate consumers' vehicle purchase decisions. Given the point estimates of our model, we find that vehicles currently rated as safe would need to collectively reduce their crashworthiness by 32% of the difference between the average safety of vehicles with and without safety badges under the current ratings scheme, which would imply that a discrete format accounts for 32% of the benefits of having ratings at all. We view this as an implausibly large impact of the discrete ratings format choice on manufacturer investments in safety.

7 Conclusion

In this paper, we investigate whether switching from the status quo discrete crashworthiness rating to a continuous crashworthiness ratings format would reduce fatalities and raise welfare. Our results suggest that such a scheme may reduced annual accident fatalities by 2,550 in the United States and raise consumer welfare in the automobile market by about 29%, compared with the discrete rating scheme observed in practice.

A remaining question for future research is whether a better discrete ratings scheme might yield even better outcomes than a continuous ratings scheme. Prior theoretical results suggest this may be possible if all vehicles have similar incentives to be rated as safe. However, there are empirical challenges. First, there is variation in the gains to improving safety across nameplates and manufacturers. Dubey and Geanakoplos (2010) notes that using multiple levels of discrete ratings may alleviate this problem, but incentives to reach

[30] The immediate impact would not be as large. The full impact would not be realized until the the existing stock of vehicles (initially purchased by consumers only with access to discrete ratings) is replaced with newer vehicles purchased by consumers that could utilize the continuous ratings in their vehicle purchase decisions.

the next threshold are reduced as the number of discrete ratings increases. Second, the theoretical benefits of a discrete rating scheme assumed agent response functions were know. In empirical applications, agent response functions are difficult to estimate (for researchers and ratings agencies) when nameplates are sold globally and subject to different crashworthiness ratings criteria in different regions. Imprecision in estimated manufacturer responses may eliminate any expected gains from a better discrete rating scheme. If the rating agency were to set the wrong threshold, either too low providing too little incentive to improve safety or too high to feasibly be achieved, firms may invest less in safety features than they would under a continuous rating scheme. The threat of accidentally setting the wrong threshold for a discrete ratings scheme may outweigh any theoretical gains achieved from a discrete ratings scheme, even with the ex-ante best discrete ratings.

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.

Berry, S., J. Levinsohn, and A. Pakes (1999). Voluntary export restraints on automobiles: Evaluating a trade policy. *American Economic Review 89*(3), 400–430.

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.

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*, 956–971.

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.

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*.

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.

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.

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.

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 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), 2421–55.

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

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

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.

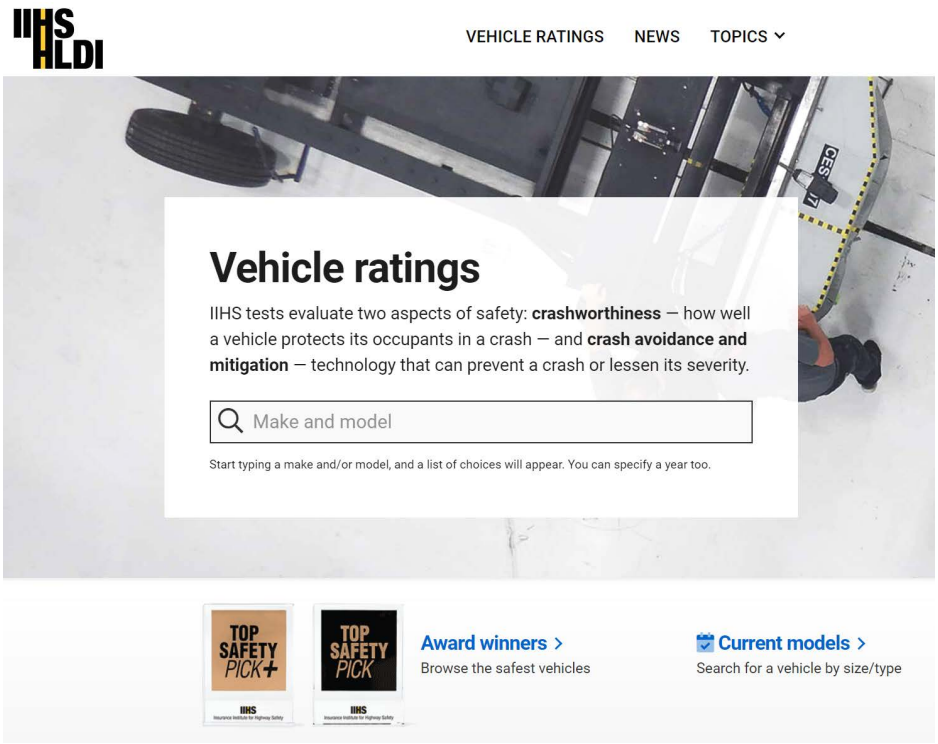

Figure 1: IIHS Ratings Landing Page

Notes: Figure shows the ratings landing page on the IIHS website. See: https://www.iihs.org/ratings. Accessed Dec 5, 2019.

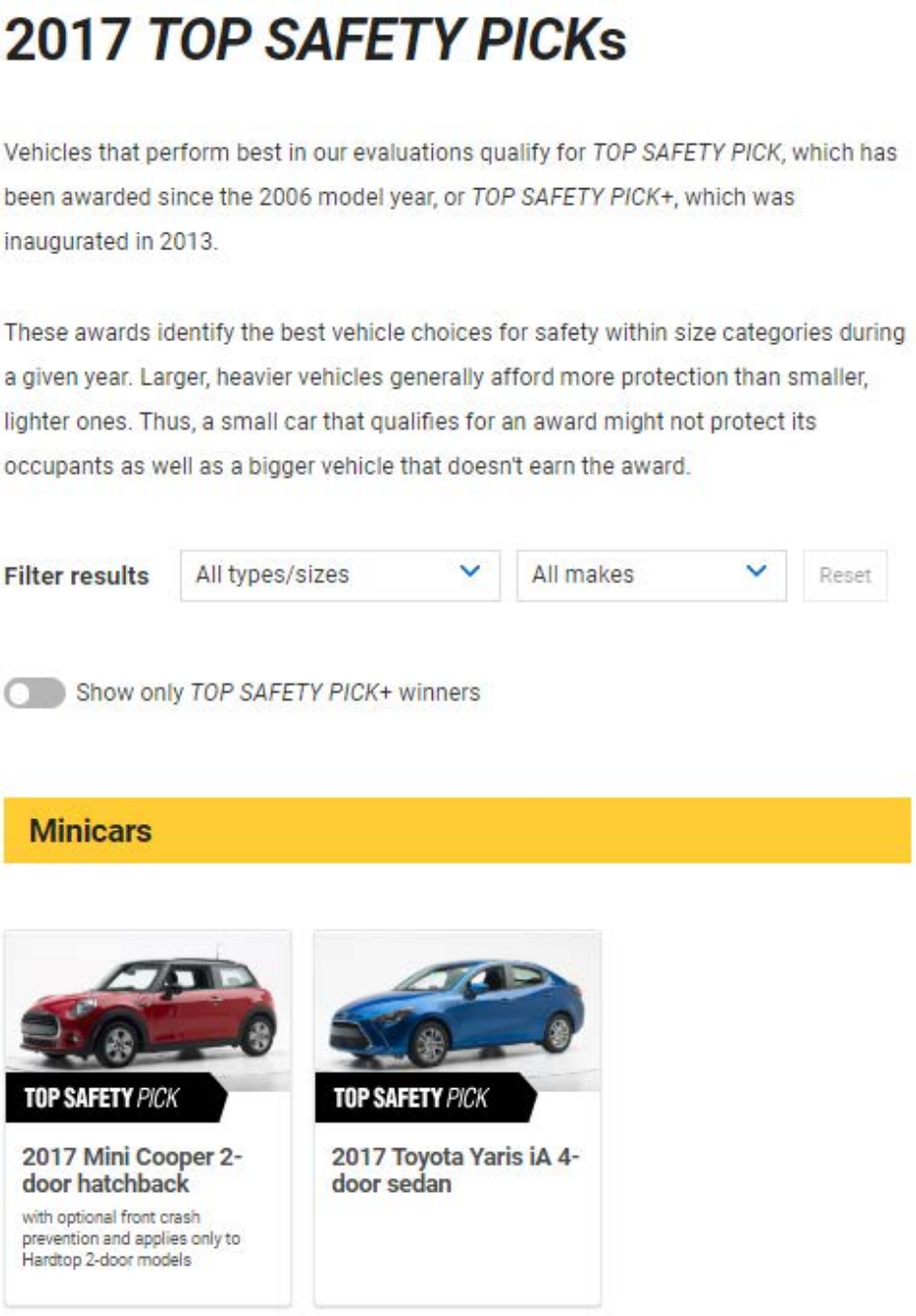

Figure 2: Top Safety Picks List

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.

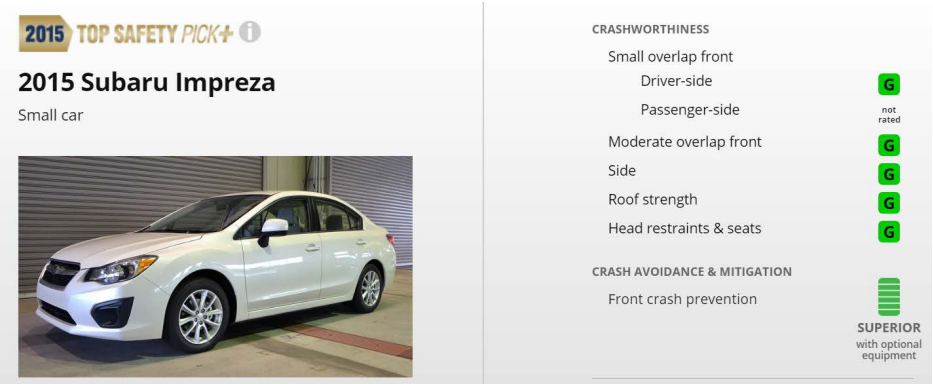

Figure 3: Example Landing Page for Specific Vehicle

Notes: Figure shows a screenshot of the 2015 Subaru Impreza rating page on IIHS's website. See: https://www.iihs.org/iihs/ratings/vehicle/v/subaru/impreza-4-door-sedan/2015. Accessed March 11, 2019.

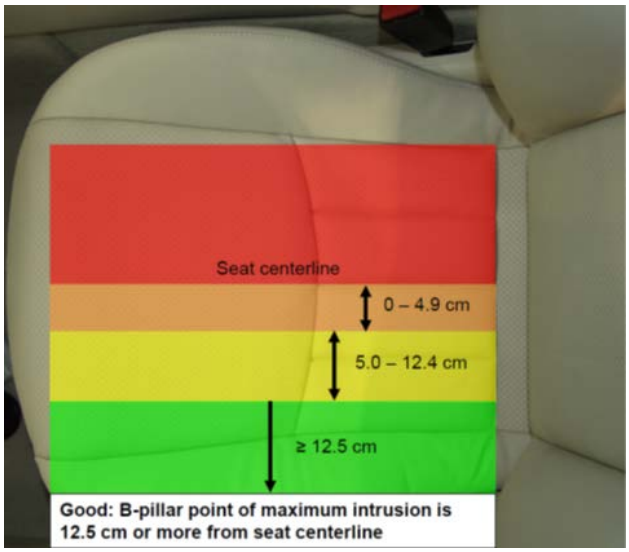

Figure 4: IIHS: Thresholds for B-Pillar Vehicle Structure Discrete Sub-Sub-Rating

Note: The B-pillar intrusion in cm 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's "IIHS Side Impact Test Program – Rating Guidelines." http://tinyurl.com/y7rg4azf

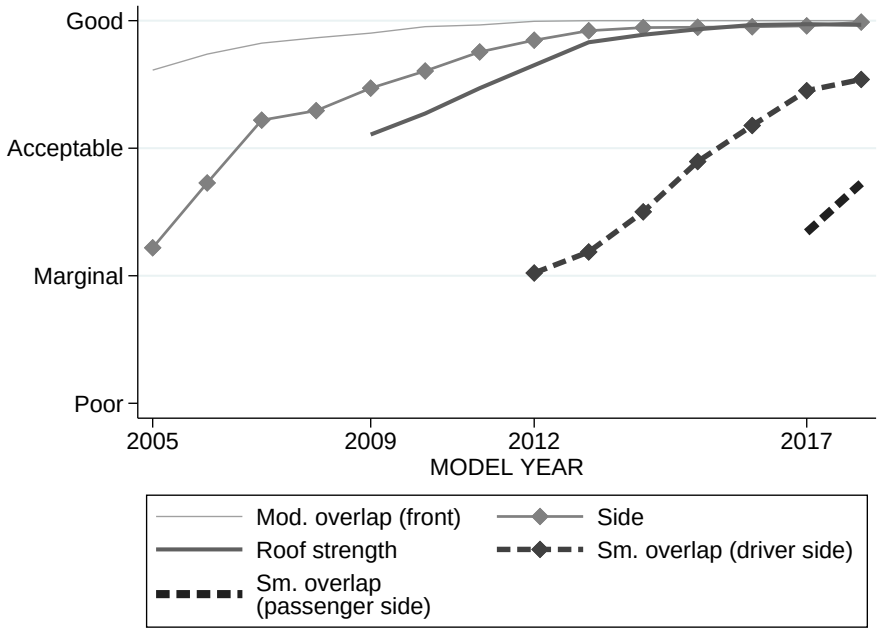

Figure 5: Evolution of Discrete Sub-Ratings [IIHS]

Notes: The figure shows average discrete ratings for each type of IIHS crash test over time, where each is rated on a scale from 1 ("poor") to 4 ("good"). Each rating is shown from its inception, or from 2005, whichever is later.

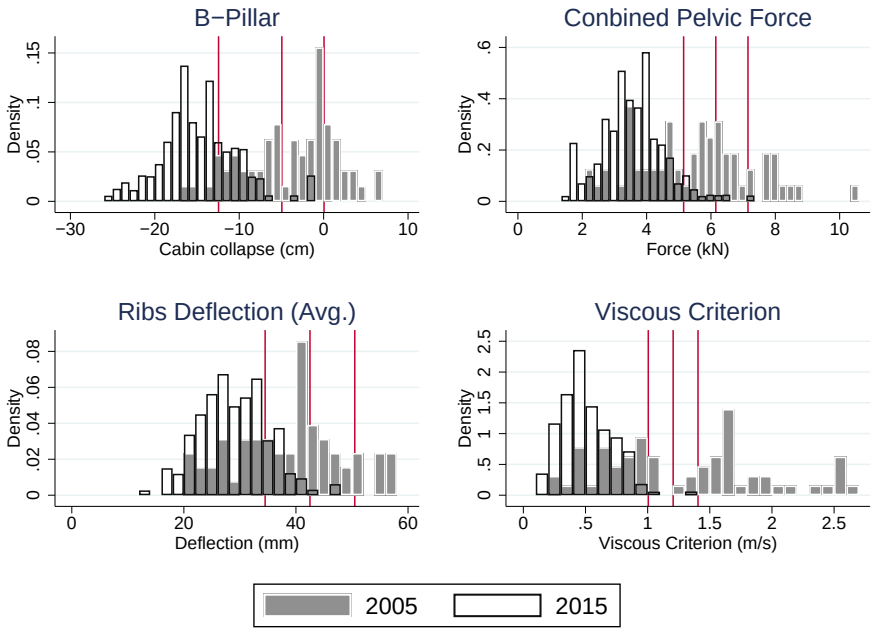

Figure 6: Evidence of Clumping at Thresholds

Notes: The figure shows histograms of the four discrete side impact crash test sub-measures which 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 a "acceptable" and "marginal" discrete sub-sub-rating. The right-most threshold denotes the cutoff between a "marginal" and "poor" discrete sub-sub-rating

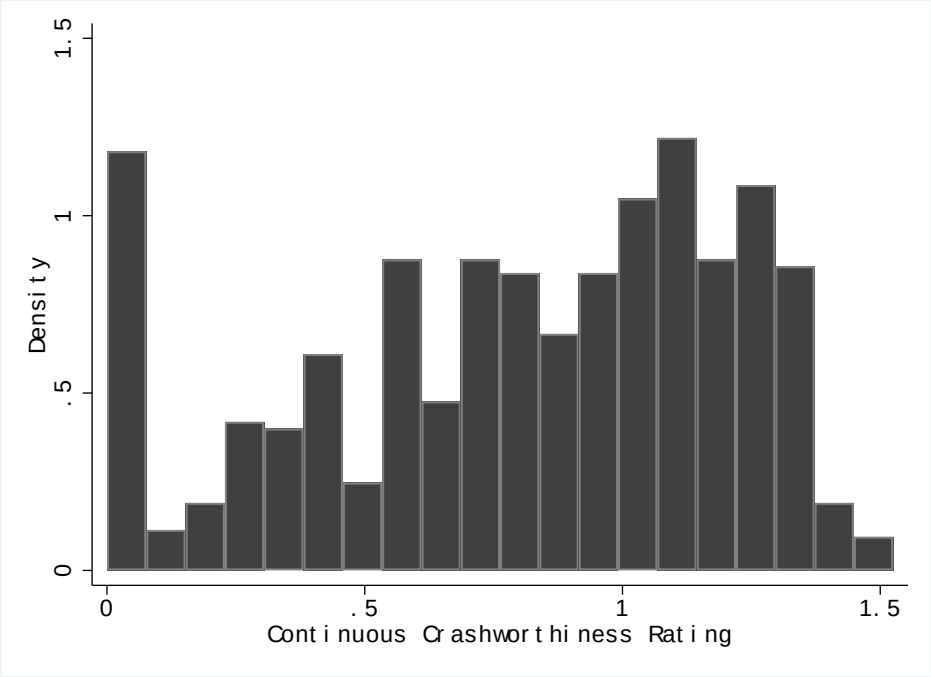

Figure 7: Distribution of Continuous Ratings Among Vehicles Earning Safety Badge

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 11.9 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 7.6 fatalities per million vehicles each quarter. Note that higher values along the horizontal axis correspond to lower driver fatality rates.

Table 1: Mapping of Sub-Ratings to Side-Impact Crashworthiness Rating [IIHS]

	Assigned demerits if sub-rating 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 sub-ratings. Sub-ratings themselves have sub-sub-ratings. For example, the torso sub-rating 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 sub-rating equals the worst of its sub-sub-ratings. If less than 6 demerits are assigned across the various sub-ratings, the overall discrete side crash rating is "good." Between 8 and 20 demerits corresponds to an "acceptable" side crash rating. Etc.

Table 2: Summary Statistics — Vehicle Characteristics

Model Year		1(IIHS Badge)	Car Size (W × L)	MPD	HP/WT	1(4WD)	Real Price (in $10,000s)
2012	Avg	0.476	1.379	0.029	0.613	0.164	1.515
	SD		0.177	0.014	0.158		0.866
2013	Avg	0.534	1.369	0.030	0.649	0.167	1.593
	SD		0.171	0.011	0.209		0.949
2014	Avg	0.238	1.368	0.034	0.659	0.157	1.661
	SD		0.169	0.018	0.224		1.009
2015	Avg	0.287	1.368	0.042	0.669	0.154	1.676
	SD		0.164	0.020	0.214		0.949
2016	Avg	0.283	1.370	0.049	0.665	0.154	1.658
	SD		0.165	0.025	0.216		0.927
2017	Avg	0.367	1.375	0.046	0.666	0.160	1.659
	SD		0.166	0.025	0.221		0.925
2018	Avg	0.260	1.387	0.042	0.669	0.190	1.709
	SD		0.161	0.019	0.197		0.932

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 3: Summary Statistics — Crash Test Measurements

Model Year		Moderate Overlap Front Crash Test							Side Crash Test			
		Structure		Head Inj.		Neck	Chest	Leg	Chest		Leg	
		Foot Intr.	A-Pillar	HIC-15	Peak G's	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	244.1	17.4	18.5	0.3	0.5	28.6	3.8	0.7	110.5
	SD	4	1	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	0.9	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 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. G's denotes g-force. Neck extension: torque in Newton Meters (Nm). N_{ij} is a linear combination of axial loads and bending moments (see *https* : *//tinyurl.com/y2sp4yot*). 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 4: Relating Fatal Accident Rates and Crash-Test Measurements

	(1)	(2)	(3)	(4)
Year Redesigned	-0.0457	-0.0130	-0.00402	
	(0.0132)	(0.0122)	(0.00960)	
I(Awarded Safety Badge)		-0.181	-0.208	
		(0.0380)	(0.0327)	
Moderate Overlap Crash Test Continuous Measures				
Footwell intr. at center (cm)		0.00949	0.0214	
		(0.00444)	(0.00408)	
A-pillar rear movement (cm)		0.0305	0.0186	
		(0.00951)	(0.0113)	
Head HIC-15		0.0000707	0.000579	
		(0.000177)	(0.000161)	
Head peak G's		0.00122	0.000285	
		(0.000460)	(0.000463)	
Neck extension (nm)		0.000662	0.00318	
		(0.00187)	(0.00172)	
Neck max N_{ij}		0.405	0.203	
		(0.183)	(0.175)	
Right tibia axial force (kN)		0.367	0.197	
		(0.0687)	(0.0709)	
Side Crash Test Continuous Measures				
Chest avg. deflection (mm)		0.0109	0.00813	
		(0.00251)	(0.00236)	
Sternum max defl. rate (ms)		0.0430	0.0451	
		(0.0121)	(0.0113)	
Left femur force (kN)		0.302	0.129	
		(0.0359)	(0.0317)	
Femur L-M moment (Nm)		0.00108	0.000828	
		(0.000252)	(0.000240)	
Vehicle age FE	Y	Y	Y	Y
Time FE	Y	Y	Y	Y
Weight/dimensions			Y	
Nameplate/design generation FE				Y
N (in millions)	2,845	2,845	2,845	2,838
pseudo R^2	0.0023	0.0063	0.0084	0.0125

Notes: The table shows results from logit models predicting the probability of driver death. Standard errors (in parentheses) are clustered by nameplate and model-year. N = the sum of (cumulative production × quarters of exposure) across vehicle nameplates. Specification 4 excludes vehicles with zero deaths. Restricting specifications 1-3 to the same sample has negligible impacts on reported outcomes.

Table 5: 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
	(0.0585)	(0.0556)	(0.0554)	(0.0508)	(0.0508)
price (α)	-1.122	-1.031	-1.024	-2.880	-2.904
	(0.181)	(0.186)	(0.185)	(1.543)	(1.550)
Size (W $\times$ 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/Weight	-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 (scale from 1 to 4)					0.0460
					(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 5 shows estimation results under various specifications of a representative agent logit model. Estimates were obtained by regressing $log(s_j) - log(s_0)$ 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 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.

Table 6: Estimated Parameters — Random Coefficient Demand Model

Variable	Parameter Estimate	Standard Error
Standard deviations (σ_β)		
1(IIHS Badge)	10.281	1.755
HP/WT	10.276	1.728
1(4WD)	5.835	5.367
Term on Price		
$1/y_i$	-12.573	5.369
Means ($\bar{\beta}$'s)		
I(IIHS Badge)	-8.548	2.251
HP/WT	-20.042	3.883
I(4WD)	-5.647	7.409
Size (W×L)	3.310	0.402
MPD	2.699	3.500
Vehicle Age FE	Y	
Vehicle Type FE	Y	
Engine Type FE	Y	

Notes: 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). The model includes a constant term. Standard errors are clustered by nameplate and model year combinations.

Table 7: Simulated Impacts of a Continuous Rating Format

	Percent Change from Status Quo
Driver Death Rate	-10.232%
	[-39.061, 0.003]
Consumer Surplus	28.913%
	[-0.208, 34.634]

Note: 95 percent confidence intervals computed via parametric bootstrapping are shown in brackets.