Do Weight-Neutral Safety Ratings Curb the Vehicle Arms Race? Evidence from Safety Award Criteria Updates

Thao Duong*

Jonathan B. Scott†

December 4, 2025

Abstract

Safety ratings are evaluated assuming equal weights between vehicles. This paper examines the tradeoffs between more accurate depictions of risk and the feedback generated when weight is accounted for. Underlying crash test outcomes are leveraged to isolate a demand response to safety ratings and to mitigate selection when identifying their impacts in traffic accidents. By incorporating effects of weight into a counterfactual rating, we demonstrate a permanent wedge would occur between current and new vehicle sizes—with a strategic response of 7 pounds per 100 pounds in a representative vehicle weight—highlighting clear benefits of neglecting weight in safety ratings.

JEL Codes: R41, L62, I18, L15

*Union College; **duongt@union.edu**

†University of Texas, Dallas; **jscott@utdallas.edu**

Any errors are our own.

1 Introduction

Evidence has shown that consumers exhibit a strong preference for large vehicles, even after controlling for characteristics like fuel economy and performance. While this demand often reflects a desire for greater capacity or utility, prior research has also highlighted the role of perceived safety in driving the shift toward heavier vehicles (e.g., Li, 2012; Scott, 2022). Large vehicles have been shown to reduce fatality risk for their own occupants in collisions, generating significant private safety benefits. However, consumers who seek protection in heavier vehicles impose substantial external costs—through higher emissions, increased road degradation, and an elevated risk to other drivers on the road.

There is a substantial literature on the external costs of vehicle weight and the consumer behavior that generates these externalities. Anderson and Auffhammer (2014) quantify the added fatality risk that heavier vehicles impose on others in two-vehicle collisions. Li (2012) examines consumer preferences for safety in vehicle choice, while Scott (2022) shows that consumers internalize vehicle size as a safety attribute by estimating how information about nearby fatal accidents affects purchasing behavior. Seminal work by White (2004) illustrates how such preferences can give rise to a vehicle "arms race," a behavioral feedback loop ultimately resulting in an inefficiently heavy—and more dangerous—vehicle fleet.

The concept of a vehicle arms race describes a prisoners' dilemma wherein consumers purchase increasingly heavy vehicles to protect themselves against other cars on the road. This behavior has clear implications for safety and produces an inefficient equilibrium, where the vehicle fleet is much heavier than it otherwise would be. How consumers best respond to the distribution of vehicle weights on the road is difficult to estimate due to the inherent endogeneity in choice. Thus, there is very limited evidence that this feedback loop exists in practice. In this paper, we explore an alternative channel to pin down how publicly available information on safety might contribute to (or mitigate) an arms race.

There are two prominent rating agencies charged with producing information on safety ratings: the National Highway Traffic Safety Administration (NHTSA) and the Insurance Institute for Highway Safety (IIHS). Ratings are derived from various in-lab crash tests, estimat-

ing the damages incurred through a controlled vehicle accident. By design, both agencies derive a final safety rating that is independent of a vehicle's size. While empirical studies have shown that heavier vehicles are objectively safer, frontal car crash tests are conducted by driving a vehicle into a fixed barrier, an outcome which simulates a two-car collision between equal sized vehicles. Thus, as currently constructed, ratings provide information on the safety of a given vehicle, within weight class, and provide little information on the role of that vehicle's weight—or other vehicle weights—in real-world accidents. This approach has clear social efficiency implications.

In this paper, we explore the potential outcome of a counterfactual methodology which would incorporate information about vehicle weights directly into the ratings. This approach arguably produces a more accurate depiction of risk, but may generate an inefficient response by consumers who subsequently wish to purchase larger cars. By estimating the response to these "full-information" safety ratings, we are able to quantify the social benefit of hiding or, "conditioning out," the role of weight in reputable sources of information on vehicle safety.

Our empirical approach is two-fold. On the front end, we estimate the consumer response to safety ratings. We construct a simple vehicle demand model, incorporating IIHS top safety picks as an additional vehicle attribute. These picks are determined based on 8 unique crash tests, the crash outcomes translated to a Likert scale describing the vehicle's performance (i.e., poor, marginal, acceptable, and good). While salient information of an IIHS safety evaluation may be correlated with other vehicle characteristics, the criteria for choosing the awards changes yearly independently of individual attributes. For example, in one year, a top pick classification may only require an "acceptable" rating for both crash test 1 and 2, while in the following year it may require a "good" classification for crash test 1. Our approach assumes that all confounding factors associated with unobserved vehicle characteristics are captured by the individual test outcomes, thus, leveraging the remaining variation in the criteria change to pin down the causal response to an IIHS award. Our approach holds constant the individual crash metrics, exploiting only their interactions with the thresholds for assignment of a top safety pick.

Of primary interest is the counterfactual demand for a vehicle under an alternative methodology which incorporates relative weight into the safety rating. To do so, we leverage empirical data on traffic accidents to recalibrate the rating. Ultimately, we explore the mechanical relationship between a top safety pick and fatality risk, then generate a mapping between fatality probabilities and the safety rating. As we are only interested in the mechanical relationship, the underlying empirical challenge is the potential selection of drivers of safer vehicles into more dangerous car crashes. To mitigate this concern, we leverage only information from the 8 individual lab test outcomes. Our approach assumes that the IIHS top pick classification is the only information salient to the consumers, while the individual test outcomes determining the safety designation are less salient. Using the unobserved (to the driver) crash test results in an instrumental variable design allows us to estimate how much safer the top picks are in traffic accidents.

The traffic fatality risk regression is both a function of the top pick designation and relative vehicle size—an additional component of risk not accounted for in the crash tests. To mitigate potential selection concerns related to vehicle size, we follow Anderson and Auffhammer (2014) and estimate the role of relative weight on the sample of two-car collisions, where the assignment of the "opposing" vehicle in the accident is assumed exogenous. Given the estimated effects of top picks and vehicle size on fatality risk, we are then able to perform a simple scaling to incorporate the effects of vehicle weight into a counterfactual safety rating.

The main objective of this paper is to recover how preferences for weight might change in response to some representative (and likely evolving) fleet weight. We formalize the thought experiment through the safety ratings channel. Our estimates suggest that the marginal effect of vehicle weight on demand would increase by 10-12 percent per 1,000 pound increase in mean fleet weight. Back-of-the envelope calculations show that this estimate implies that consumers would respond to a 100 pound increase in fleet weight by increasing their vehicle size by 7 pounds. Under limited turnover of vehicles on the road, we demonstrate that, in equilibrium, the ratings change will introduce a permanent wedge between new vehicle weights and fleet-wide averages, triggering an immediate external cost of up to $215 million in traffic fatalities. These costs are distributional in the sense that new, heavier vehicles

are always safest while imposing the most harm to older vehicles on the road. Given the spiraling growth in vehicle sizes under the hypothesized safety ratings — and the significant external costs of large vehicles in terms of safety, environmental, and infrastructure impacts — our findings highlight the tradeoffs ratings agencies face when seeking to provide more accurate depictions of risk.

The results in this paper align with a broader literature showing that consumers respond strongly to salient but incomplete signals. A large literature documents myopia or incomplete internalization of energy attributes, especially fuel economy (Busse, Knittel, and Zettelmeyer, 2013; Allcott, 2013; Allcott and Wozny, 2014; Sallee, West, and Fan, 2016).[1] Simple regulatory incentives also move demand (Hoekstra, Puller, and West, 2017), and rational inattention can rationalize undervaluation of efficiency (Sallee, 2014). Salient safety messages can even backfire; for example, public display of recent accident statistics have been shown to raise crash rates (Hall and Madsen, 2022). These patterns motivate why prominent safety designations may disproportionately shape choices even if they imperfectly map to true crash risk. A large body of work also shows that quality disclosure through ratings, labels, or certification routinely shifts demand (e.g., through restaurant grades (Jin and Leslie, 2003), online reviews (Luca, 2011), environmental labels (Houde, 2014, 2018), health claims (Ippolito and Mathios, 1990), and school report cards (Hastings and Weinstein, 2008)).

It has been well-established that greater vehicle mass improves own-occupant safety (Crandall and Graham, 1989; Van Auken and Zellner, 2005; Anderson, 2008), while heavier vehicles impose additional risks to others on the road. It has been estimated that 1,000 lb increase in vehicle mass can raise opposing vehicle fatality risk by up to 46–47% (Jacobsen, 2013; Anderson and Auffhammer, 2014). This relationship exists even when the weight difference comes in the form of an additional passenger (Evans, 2001). This introduces a tradeoff: when safety ratings emphasize the objective role of weight, they may inadvertently encourage choices that improve private safety while worsening societal outcomes.

Consequently, information is not always welfare-enhancing when externalities are present.

[1] See also Allcott (2013) on MPG vs. cost-per-mile perceptions.

Public disclosure can reduce welfare (Morris and Shin, 2002), optimal policy may involve coarse or partial signals (Kamenica, 2019), and in some cases, restricting information can improve outcomes (Arato and Nakamura, 2022). Safety interventions can also trigger risk compensation (Peltzman, 1975; Peterson, Hoffer, and Millner, 1995; Cohen and Dehejia, 2004), with net-negative effects. Our setting reflects these dynamics: when weight is introduced as a salient safety attribute, consumers substitute toward heavier vehicles, producing a ratings-induced arms race with an inefficient equilibrium. To our knowledge, this is the first paper to isolate this mechanism within safety ratings. By holding the direct response to weight constant, our counterfactuals show that periodic rating updates disproportionately reward the heaviest vehicles, generating a feedback loop in vehicle size. These results suggest that limiting information on weight—a current practice—may help dampen inefficient substitution and improve overall welfare.

The paper proceeds as follows. Section 2 outlines institutional background on safety awards and describes the datasets on ratings, vehicle sales, and traffic accidents. Section 3 details the empirical strategy, including construction of the counterfactual (mass-adjusted) rating and the demand and traffic fatality risk specifications. Section 4 presents the main estimates and safety rating counterfactual calculations. Section 5 concludes the paper with a brief discussion of the implications of our results.

2 Data and Institutional Setting

This section provides background on safety ratings methodologies and describes the datasets used; mainly, IIHS rating information, VIN-decoded new-vehicle sales and characteristics, and police-reported traffic crashes and fatalities. It concludes with a summary of the key variables used in the analysis.

2.1 Safety Ratings

There are two primary vehicle characteristics of interest in this paper: weight and safety ratings. These attributes are used both in a demand framework and to estimate their role in traffic fatalities. This section discusses the safety ratings used in the paper.

Historically, two prominent ratings agencies represent the dominant sources of information on vehicle safety; each agency deriving their safety ratings through in-lab crash tests. The National Highway Traffic Safety Administration (NHTSA) conducts periodic safety testing on various models and publishes their ratings on their website.[2] There are several limitations to these data, however. NHTSA crash tests are not universally conducted across all vehicle models due to excessive costs. Further, the vehicle models which do have ratings available are only tested on particular model years. This leaves many gaps in the data in which a safety rating for a given make-model-year may not be observed. Given these limitations, we conduct our analysis using crash ratings published by the Insurance Institute for Highway Safety (IIHS), whose breadth and frequency of testing generally surpasses that of NHTSA. In Appendix A, we attempt to overcome some of the issues with the NHTSA data and provide some estimates as a basis for comparison.

There are many similarities in the testing methodologies between IIHS and NHTSA. For example, each agency produces their ratings on a scale which may only be interpreted within a given weight class. On their website, NHTSA offers the disclaimer: "Overall Vehicle Scores can only be compared to other vehicles in the same class and whose weight is plus or minus 250 pounds of the vehicle being rated." The reason is due to the controlled design of the crash tests, which explicitly hold weight constant. IIHS describes their frontal crash tests as one that simulates a collision between two vehicles of equal size: "The forces in the test are similar to those that would result from a frontal offset crash between two vehicles of the same weight, each going just under 40 mph." These "weightless" safety ratings motivate our research question.

The IIHS was established in 1959 by three leading insurance associations that collectively represented 80 percent of the U.S. auto insurance market. The nonprofit agency produces annual publications, awarding the best performing vehicles IIHS "top" and "top plus" safety picks. The information is widely publicized, often cited in manufacturer marketing campaigns when awarded top pick status. Still, we acknowledge that this is only one source of information on safety available to consumers and, thus, the results of this paper should be in-

[2] nhtsa.gov/ratings

terpreted through the lens of IIHS ratings, rather than a combination of all available sources.

The IIHS awards are the result of several tests aimed at estimating a vehicle's performance in a car crash. For each of the 8 tests, IIHS records specific injury and structural measurements (e.g., HIC-15, Nij, chest deflection, tibia index, intrusion) which are subsequently mapped into a categorical Likert rating: the ratings bands. Tests 1-7 are evaluated on the four categories: "poor", "marginal", "acceptable", "good". The frontal crash prevention technology is evaluated in test 8 on, first, if its availability and, second, whether the equipped technology is classified as "basic", "advanced", or "superior" according to IIHS metrics. Importantly, the rating bands are set within each test's rating protocol and it is possible that IIHS may update those protocols in any given year. While the cutoffs for the ratings bands may change, identification in this paper holds the ratings bands independently constant (e.g., comparing two vehicles who score "Good" on test 1 in year t), while leveraging the criteria changes each model year, which combine all 8 test outcomes into a single award indicator. The thresholds for each of the 8 tests in determining a top pick are revised each year, generating variation in criteria across model years.

We use the agency's public API to gather information for each crash test and merge the data by model to our sales and traffic accidents data. Both the individual crash tests and the ultimate top pick designations are essential components of our empirical design. The designation criteria in each model year in our data set are illustrated in Table 1. The identifying variation comes from the interaction of the results from the 8 primary tests, and the changes in the thresholds which determine a vehicle's top pick classification.

2.2 Vehicle Sales and Characteristics

New vehicle sales make up our main outcome of interest in estimating a vehicle choice model. The data derive from Texas vehicle registrations from the Department of Motor Vehicles (DMV), and are reported at the county-month level for the years 2014-2019. For our analysis, we aggregate sales to the metropolitan statistical area (MSA) level. The choice of MSA-level sales is chosen as a better representative sample of a market for vehicles; we

observe multiple sales for the vast majority of vehicle models in our sample at this level of aggregation. Registrations of new vehicles are reported by their unique 17-digit vehicle identification number (VIN). To obtain the specific characteristics of each vehicle, we make use of a VIN decoder.

Vehicle characteristics come from DataOne Software, in which we directly join in unique information on each vehicle based on the 10-digit VIN stub (VIN10). A VIN10 defines a vehicle up to their make-model-year-trim level, in addition to particlar packages. Our analysis aggregates vehicles into a make-by-model-by-year-by-fuel type index, as occasionally, a model will be produced in multiple fuel types. The characteristics of interest are generally constant within this narrow vehicle description. The main characteristic of interest in this paper is a vehicle's weight, as defined by its curb weight.[3]

Once all characteristics are collected, we combine information on reported initial registration dates and model year in order to infer new vehicle sales. For this, we assume that a sale that takes place in a year in or preceding that of the model year is a new vehicle sale.

2.3 Traffic Accidents

A recalibration of the safety ratings to one which accounts for vehicle weight is of primary interest in this paper. Given the infeasibility of redesigning the crash tests to account for vehicle relative weights in two-car collisions, we gather this information directly from empirical data on traffic accidents.

Data on traffic accidents are collected from the Texas Department of Transportation's (TxDOT) Crash Records Information System (CRIS). CRIS is a comprehensive source of all reported accidents in Texas, with granular information describing the characteristics of an accidents at the crash, vehicle, and person level. The main analysis is conducted on the sample of two-car collisions, in which the VINs of the vehicles involved are the primary identifiers and allow us to merge together relevant vehicle characteristics.

The primary outcome of interest is a traffic fatality occuring, defined at the vehicle level.

[3]The curb weight of a vehicle accounts for all standard components, including a full tank of gas and all necessary operating fluids.

The VIN allows us to merge in any relevant characteristics (including safety ratings). Other characteristics of the crash such as county, date of crash, and speed limit offer additional controls.

2.4 Summary of Key Variables

A summary of key vehicle characteristics appears in Table 2. The unit of observation is a unique model; we report unweighted means and standard deviations separately for vehicles awarded an IIHS Top Safety Pick, those that are not, and for the full sample. Although these differences should not be interpreted as a direct relationship, non-Top Safety Pick models are roughly 300 pounds heavier on average. While several factors could drive this pattern, one would expect the opposite if ratings were mechanically a function of mass. Results in Section 4 confirm that rating assignments are independent of weight after accounting for model-specific confounders.

Table 2 also reports crash counts and traffic fatalities from the accidents data. Both measures are higher among non–Top Safety Pick models, though these differences should not be interpreted as causal due to potential selection. Our empirical strategy presented in Table 3 aims to mitigate these concerns.

Because the marginal effect of weight on Top Safety Pick assignments is effectively zero by construction—reflecting the physical design of the crash tests—a primary objective of this paper is to construct a counterfactual rating that fully incorporates information on vehicle mass, producing an, arguably, more accurate depiction of risk. As an illustration, Table # lists, by model year, the heaviest models that did not receive a Top Safety Pick and the lightest models that did receive the award. These models illustrate the most likely to be affected by weight-adjusted ratings; in particular, they are the most likely to have their Top Safety Pick status "flip" once weight is incorporated.

3 Empirical Strategy

This paper seeks to measure the demand responses to a counterfactual ratings methodology which incorporates the role that weight plays in vehicle safety. We hypothesize that when weight is indirectly internalized through this additional channel, a feedback loop is generated whereby consumers purchase increasingly heavy vehicles; particularly when *relative* weight on the road is of significance.

Our empirical methodology consists of two primary components. First we estimate a simple vehicle demand model, identifying the causal response to IIHS safety ratings. Next, we recalibrate the historical ratings by incorporating vehicle weight. This requires us to create a mapping between the crash test data and empirical fatality risk, combining the estimated effects of weight on crash outcomes.

3.1 Demand Model

We begin with a standard logit model of vehicle choice, which includes the IIHS top safety pick designation as an additional vehicle attribute. For a vehicle model j in model year t, define the variable top_{jt} as a binary indicator representing the top safety pick designation. The primary empirical challenge in estimating a causal response to safety pick assignment is that the general safety of a vehicle is likely correlated with other, unobserved characteristics. These factors may upward bias our estimates—for example, if they are correlated with vehicle quality—or downward bias—for example, if safer vehicles are less stylistically appealing. Our objective is to leverage IIHS critera changes for top safety pick awards in order to isolate the actual information treatment of the assignment.

For an individual crash test indexed by $r = 1, ..., 8$, we can describe the ratings bands, or crash test outcomes on the set $\{Poor, Marginal, Acceptable, Good\}_{r=1,...,7}$, for tests 1-7, and the set $\{Not\ Available, Basic, Advanced, Superior\}_{r=8}$, for test 8. The test performance for each vehicle can then be described by the vector of dummy variables $R_{rjt} = (\mathbb{1}^{poor}_{rjt}, \mathbb{1}^{marg}_{rjt}, \mathbb{1}^{acc}_{rjt}, \mathbb{1}^{good}_{rjt})$, or $R_{8jt} = (\mathbb{1}^{na}_{8jt}, \mathbb{1}^{basic}_{8jt}, \mathbb{1}^{adv}_{8jt}, \mathbb{1}^{sup}_{8jt})$ for test 8, where $\mathbb{1}^{k}_{rjt}$ indicates whether test r produces an outcome of *at least* a k classification for model j in year t. Let the full vector of dummy

variables across all 8 tests be defined as $R_{jt} = (R_{1jt}, ..., R_{8jt})$. This describes a vehicle's overall test performance and completely determines its top pick designation, conditional on the criteria in each model year. While changes in testing protocols might alter the way that the ratings bands, R_{rjt}, are assigned in any given year, the objective is to ultimately isolate changes in the award criteria, holding the ratings bands for a particular year independently constant.

Table 1 describes the relative outcome required in each test to be awarded a top safety pick. For example, in model year 2018, a top safety pick is defined as $top_{j,2018} = \mathbb{1}^{good}_{1j,2018} \times \mathbb{1}^{good}_{3j,2018} \times \mathbb{1}^{good}_{4j,2018} \times \mathbb{1}^{good}_{5j,2018} \times \mathbb{1}^{good}_{6j,2018} \times \mathbb{1}^{acc}_{7j,2018} \times \mathbb{1}^{adv}_{8j,2018}$. Notice how crash test 2 is not relevant in 2018 and, thus, only the minimum requirement (i.e., poor) was required. In contrast to the 2018 criteria, there was no minimum requirement for crash test 7 in 2017 and, in 2019, a top safety pick required at least an "acceptable" rating for crash test 2. In a general sense, we can define a top safety pick award for some vehicle j in year t given a known criteria function, $T^t(\cdot)$. That is,

$$top_{jt} = T^t(R_{jt}) \tag{1}$$

Intuitively, we should be able to leverage the changes in the award criteria while holding the underlying test outcomes constant. For example, key variation of interest should reside in the difference $T^t(R_{jt}) - T^{t-1}(R_{jt})$, to exploit the criteria change over model years, holding constant test performance for that year, R_{jt}. Our approach proceeds with the spirit of this idea. In base specifications, we control for the lagged criteria function directly, $T^{t-1}(R_{jt})$, as a counterfactual rating, and leverage only the deviation in criteria for year t. As R_{jt} is simply a vector of discrete outcomes, in our main specifications, we simply control for these dummies directly. Specifically, we estimate the following equations.

$$\begin{aligned} log(q_{cjt}) &= \beta top_{jt} + \rho T^{t-1}(R_{jt}) + \phi_j + \lambda_{cm_jt} + \varepsilon_{cjt} \\ \text{or} \qquad &= \beta top_{jt} + R_{jt}\Gamma + \phi_j + \lambda_{cm_jt} + \varepsilon_{cjt} \end{aligned} \tag{2}$$

where the outcome of interest is logged, aggregate quantities of model j in model year t,

purchased in city c. Aggregate city-year purchases in the denominator of market shares are absorbed into city-year fixed effects and, thus, our specification is a standard logit model. As manufacturers have the potential to lobby for IIHS criteria changes, we control for city-by-year effects at the make of model j, m_j, level, λ_{cm_jt}. We further control for model level effects in ϕ_j. The marginal effect of interest is the coefficient β, estimating a causal response to the award. In the first equality in Equation 2, we follow our original logic, leveraging lagged criteria directly as a counterfactual. In the second equation, we introduce granular controls for each performance outcome.

In some specifications, we include additional characteristics, including vehicle weight. Vehicle price, proxied by MSRP, is introduced, but we do not interpret it directly due to endogeneity concerns. While we do not observe a significant effect of MSRP on our primary estimate, prices are held out of our main specification in order to properly identify the overall, reduced form effect of the safety ratings. It's important to note that IIHS awards are given after manufacturers have published the MSRPs for their fleet.[4] However, if manufacturers could respond to higher demand for vehicles with top safety picks by increasing prices, in order to measure the full, reduced-form effect of the top safety pick, it would be important to exclude MSRP from our regressions.

A recalibration of the safety ratings to one which accounts for vehicle weight is of primary interest in this paper. As an illustration, Table # gathers the top 5 heaviest models in our data that were not awarded a top safety pick and the top 5 lightest models who were awarded a top safety pick in any given model year.

3.2 Empirical Fatality Risk

The second component of the empirical strategy involves a recalibration of the top safety pick assignments (i.e., top_{jt}) to incorporate additional information on vehicle weight. As frontal crash rating tests simulate a two-car collision of equal sized vehicles, the marginal effect of weight on the ratings is zero by design. Therefore, to recalibrate the ratings, we

[4]There are several sources stating the timing of IIHS top safety pick awards, which after MSRPs are already published. For example, https://www.globenewswire.com/news-release/2025/03/13/3041862/0/en/IIHS-unveils-2025-TOP-SAFETY-PICK-award-winners.html?

must leverage empirical data on traffic accidents.

We begin with a simple linear probability model depicting fatality risk. The model is estimated on the subset of two-car collisions, where vehicle weights are included in a manner similar to Anderson and Auffhammer (2014).

$$fatality_{ijt} = \alpha + \delta top_{jt} + \gamma(weight_{ijt} - weight_{it}^{opp}) + \varepsilon_{ijt} \tag{3}$$

where the outcome indicates the binary outcome of a death occuring from the collision. δ describe the mechanical relationship between the safety rating and fatality risk when drivers and cars are randomly assigned to accidents and γ is the marginal effect of relative weight, where $weight_{ijt}$ describes driver i's vehicle weight and $weight_{it}^{opp}$ is the opposing vehicle weight. For simplicity, j is collapsed to an index describing make, model, and year of the vehicle, while t is redefined as the year of the accident.

In Equation 3 and throughout the paper, we assume symmetric effects of vehicle weight on fatality risk. This assumption follows from conservation of energy in elastic collisions (Landau and Lifshitz, 1976), where changes in mass affect energy distribution symmetrically. The energy output from collisions between two objects can be described by standard, nonlinear elastic collision equations (specifically, the one-dimensional elastic collision equations). In this paper, we assume constant marginal effects for simplicity and interpretation.

When δ is identified as the reduced fatality risk of a top pick, we can directly recalibrate the safety pick designations accounting for the role of vehicle weight. The procedure simply re-scales fatality risk to match top pick probability, implicitly assuming a linear mapping between the two metrics. The counterfactual safety ratings are then calculated as the following.

$$top_{jt}^{*} = top_{jt} + \frac{\partial top}{\partial w}(w_{jt} - \bar{w}) = top_{jt} + \frac{\gamma}{\delta}(w_{jt} - \bar{w}) \tag{4}$$

where $\bar{w}$ defines a baseline vehicle weight and w_{jt} is the vehicle's weight. When $w_{jt} = \bar{w}$—as is imposed by design in the controlled frontal crash tests—the counterfactual rating equals

the true rating. Vehicle safety derives from various other technologies besides weight, and those factors are maintained in the baseline rating. Two-car collisions are only one of many other potential crash outcomes, thus, the degree to which the top rating is relevant to these specific accidents, as measured by δ, enables us to scale the effects of weight appropriately. Given that the rating agency might update mean fleet weights, $\bar{w}$, on a regular basis, one can see how a feedback loop might be generated under this counterfactual safety evaluation.

As drivers and vehicles are not randomly assigned to accidents, the mechanical relationship describing the empirical safety of a top pick is not identified through direct estimation of Equation 3. In addition to the direct effects of safety, there is likely a behavioral component captured; for example, if drivers of top safety picks drive more recklessly (e.g., Peltzman, 1975). The objective is to separate this behavioral response from the mechanical relationship. To explain the intuition behind our strategy, let the following equation illustrate a linear projection of the top safety pick rating on the individual crash test dummies.

$$top_{jt} = R_{jt}\tilde{\Gamma} + u_{jt} \tag{5}$$

u_{jt} contains valuable information on the relevance of each element of R_{jt} in the top pick designation, a classification which is modified by IIHS periodically. Thus, while the outcome of each individual crash test (i.e., R_{jt}) may not be salient to the consumer, the top pick is, resulting in potential selection into fatal accidents based on u_{jt}. Controlling for u separately in Equation 3 resembles a control function approach, explicitly accounting for the behavioral response, further isolating the mechanical effect of safety ratings. This is the intuition behind our strategy and, as the control function approach is equivalent to two-stage least squares, we implement an instrumental variables design, leveraging the individual crash test outcomes to pin down the mechanical relationship between IIHS safety ratings and empirical fatality risk.

Finally, given the potential for consumers to additionally select into fatal car crashes based on their vehicle size, we will only interpret our estimate of relative weight, γ, based on independent variation in opposing vehicle weights. As vehicles in two-car collisions are presumed to be matched together in a plausibly exogenous manner, this approach should credibly pin

down the causal role of vehicle size in fatality risk. We additionally include varying sets of geographic and vehicle-specific controls and fixed effects to further isolate this variation.

3.3 Effect of Weight-Incorporated Safety Ratings

The previous subsection describes estimation of a linear fatality risk model and subsequent re-calibration of safety ratings. Ultimately, our interest is in the counterfactual demand response to this new rating (defined previously by Equation 4), and the higher, *implied*, preference placed on vehicle weight. However, in such a linear framework, preferences for weight do not vary by average weight and product-independent, additive changes in mean-weight are irrelevant to choice (i.e., $\bar{w}$ drops out). Thus, the presence of an arms race depends on the degree of strategic complementarity in vehicle weight preferences. Mathematically, we need to show that $\frac{\partial^2 top_j^*}{\partial w_j \partial \bar{w}} > 0$. To allow for these interaction effects, we impose additional distributional assumptions on unobserved fatality risk and estimate the inherent nonlinearity in death probabilities by means of probit regression. We estimate the following equation.

$$\mathbb{E}[fatality_{ijt}|top_{jt}, weight_{ijt}, weight_{it}^{opp}] = \Phi\big(\alpha + \delta top_{jt} + \gamma(weight_{ijt} - weight_{it}^{opp})\big) \tag{6}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function. This model is estimated by instrumental variables probit, using the same instruments outlined previously. Under this framework, we modify the counterfactual top safety pick ratings to account for additional nonlinear effects of vehicle weight. This becomes the following:

$$top_{jt}^* = top_{jt} + \frac{\Phi\big(\alpha + \gamma(w_{jt} - \bar{w})\big) - \Phi\big(\alpha\big)}{\Phi\big(\alpha + \delta\big) - \Phi\big(\alpha\big)} \tag{7}$$

Equation 7 is a generalization of Equation 4 under the revised probit framework. The numerator of the right-hand-side adjustment term estimates fatality risk at $top_{jt} = 0$ and subtracts the constant, $\Phi\big(\alpha\big)$, as a normalization such that $top_{jt}^* = top_{jt}$ when $w_{jt} = \bar{w}$. The denominator measures the marginal effect of a top safety pick at zero weight differential. This properly scales fatality risk values to top safety pick levels, in a manner similar to Equation 4. Under these modeling assumptions, we empirically show that this formulation gives us

strategic complementarity under the estimated parameters.

The linear estimates from Section 3.2, along with the demand parameters, give us the combined coefficient of interest, $\beta\gamma/\delta$, estimating the additional marginal effect of weight through the top safety pick channel. We can compare this directly to the average marginal effects of weight from our nonlinear specification. Of particular interest, however, is the strategic best responses inherent in an arms race. Thus, we provide estimates for the interaction effect of average vehicle weights, derived from our nonlinear model, in Section 4.[5]

4 Results

In this section, we present the main parameter estimates from both our demand and fatality risk models. Finally, the estimates of interest combine the two models and explores various counterfactual estimates of a new ratings regime.

4.1 Demand Estimates

The main estimates from our demand model are presented in Table 4. The coefficient of interest is that on the top safety pick indicator: β in Equation 2. Column 1 reports a standard three-way fixed effects specification, which controls for model, city, and make-year effects. This does not control for other factors related to information of the safety award, such characteristics inherent to vehicle safety. Column 2 leverages year-to-year deviations in IIHS criteria changes, in an effort to hold constant other vehicle attributes, thereby isolating the information shock. Column 3 reduces the three-way fixed effects to a two-way fixed effects specification by including city-make-year fixed effects. Column 4 introduces granular dummy variables for each test peformance outcome; i.e., R_{jt}. The lagged pick is removed in Column 5 and additional vehicle characteristics are added in Column 6.

Specifications 5 and 6 are our preferred specifications, as this approach does a better job specifically isolating the criteria changes away from other attributes that are correlated with a vehicle's inherent safety, as measured by its test performance. Estimates in Column 6 demonstrate the independent role of size — as measured by curb weight — on vehicle de-

[5] Appendix B directly estimates an interaction between weight by linear regression, obtaining similar results.

mand. While the estimates indicate that consumers generally prefer larger vehicles, the absence of an effect on the Top Safety Pick estimate confirm the ratings' independence from factors related to mass.

We caution against direct interpretation of the MSRP coefficient in Column 6 given the likelihood of price endogeneity, as this specification simply serves as a test for robustness. While the total effect of the safety ratings on sales are of primary interst — e.g., not holding price constant — it's important to note that the safety pick awards are made after MSRPs have been published.[6] This is likely the primary reason we do not observe a significant change in our primary estimate.

Estimates from the preferred specifications suggest that a top safety pick award is associated with a 11 percent increase in the demand for that model.[7] Given an average market share of 0.6 percent for a particular model, this would suggest a top safety pick increases market shares to approximately 0.66 percent on average. In Appendix A, we conduct a similar strategy on a subsample of vehicles with available NHTSA safety ratings.[8] Estimates from this exercise prove both qualitatively and quantitatively similar to the ones reported in Table 4. Worth noting is that, while these estimates reinforce existing evidence that consumers place significant value on vehicle safety (Alberini and Austin, 1999; Li, 2012; Jacobsen, 2013; Greenstein and Huang, 2017; Scott, 2022), this is likely not a pure consumer response, as promotional activities by dealers and manufacturers are endogenous to the top safety pick announcement. However, interest ultimately lies in the overall effect of the award on sales and subsequent fleet composition.

Figure 1 presents a falsification test for our estimates. The figure is analogous to an event study specification. The plotted coefficient estimates are derived from the following regres-

[6]For example, https://www.globenewswire.com/news-release/2025/03/13/3041862/0/en/IIHS-unveils-2025-TOP-SAFETY-PICK-award-winners.html?

[7]Estimates from a simple regression that maintains the test performance dummies but omits fixed effects and vehicle characteristics produces similar results, though are not statistically significant.

[8]Not reported here, we also test for an interaction effect between curb weight and the safety rating. This allows us to test whether ratings are less relevant for larger vehicles—e.g., consumers substitute weight for safety ratings. Estimates are statistically and economically insignificant, with estimates suggesting a response to a top safety pick shrinking by 0.1 percent per 1,000 pounds.

sion.

$$log(q_{cjt}) = \sum_{s=-\underline{L}}^{\bar{L}} \beta_s T^{t+s}(R_{jt}) + R_{jt}\Gamma + \phi_j + \lambda_{cm_jt} + \varepsilon_{cjt} \tag{8}$$

Equation 8 holds constant the test performance, R_{jt}, in year t, and evaluates it according to the criteria thresholds in multiple placebo years. As the criteria generally become more stringent over time, older thresholds — or lagged terms — should be less relevant than current or even future criteria. Future criteria normally encompass the criteria for past qualifications — i.e., a vehicle that satisfies the criteria in future years is normally a top qualifier in the current year — so it would be unsurprising to see significant effects on the leading terms.[9]

The number of lagged criteria functions is $\underline{L}$ — we choose $\underline{L} = 4$ — and the number of leading criteria functions is $\bar{L}$ — we choose $\bar{L} = 1$. We are generally able to evaluate the criteria function on R_{jt} outside of our sample, however, are limited in variation in year-to-year criteria changes. For example, the requirements for top safety awards are identical in 2014 (out-of-sample) and 2015 (in-sample), and also between 2020 and 2021 (both out of sample). This does not inhibit estimation but only limits the degree of variation in the changes in thresholds from year to year. This may explain the lack of precision in our estimates. However, Figure 1 genenerally illustrates what would be expected: earlier criteria have no economically significant effects on current demand, as positive responses become apparent on the year.

Specifically, this specification holds constant the test performance in year t, R_{jt}, estimating the effects of lagged and leading criteria changes. The estimates are normalized to the effect two years prior to the current criteria. These estimates leverage the 5 years of data in our study sample and report insignificant effects in the placebo years and a significant effect for the contemporaneous criteria; although these estimates are larger than our main estimates, likely due to changes in reference year.

[9] A model in year t can never be a top qualifier for criteria in $t+1$ if a specific technology and test did not exist at t that was required for the award in $t+1$.

4.2 Fatality Risk Mapping

Table 5 reports the estimates from the linear fatality risk model in Equation 3. The coefficients of interest are on Top Pick (divided by 100 for clarity of estimates) and vehicle weights (in 1,000s, pounds). Column 1 presents a naive specification which aims to directly measure the empirical safety of an IIHS top pick. We control directly for vehicle type, model year, and crash year-by-county fixed effects. While the negative sign indicating a drop in fatality risk is expected, we also expect an upward biased estimate when top safety pick drivers are more likely to select into high risk accidents. Column 2 introduces the vehicle weights and additional attributes as controls, which do not seem to significantly correct the estimate on Top Pick. This is unsurprising if we expect the primary bias to arise through the selection mechanism.

Two-stage least squares estimates are reported in Columns 3-4, which leverage the test performance dummies as instruments, as outlined in Section 3. Column 3 is our preferred specification. While estimates on own-curb weight and opposing weight are similar in absolute value, we interpret only the coefficient on opposing vehicle weight as the identified parameter on weight differential due to potential selection concerns. Columns 4-5 are provided for robustness and include additional vehicle attributes. Column 5 collapses vehicle type and model year fixed effects into type-by-year effects.

The estimates of interest do not change significantly across Columns 3-5, each suggesting a 0.18 percentage point reduced fatality risk for vehicles classified as a top safety pick. These estimates are marginally larger than the naive estimates, suggesting approximately 0.07 percentage points attributed to potential selection. In percentage terms, given only a 0.11 baseline fatality rate in our sample of car accidents, these estimates suggests that top safety picks are associated with a meaningful 160 percent reduction in fatality probability.

The estimates on opposing vehicle weight suggest a marginal effect 0.034 percentage points per 1,000 pounds, or a 31 percentage change in baseline fatality rate. These latter estimates can be compared to results documented by Anderson and Auffhammer (2014) who show a 1,000-pound increase in vehicle weight raises the baseline fatality probability by 47

percent. While our estimates are in the general ballpark of prior estimates, the relatively smaller effects may be attributed to newer data and safer vehicles. For example, we only use crash data for model years 2015-2019, while the average model year used by Anderson and Auffhammer (2014) is 1992.

Our nonlinear, IV-probit estimates for Equation 6 are reported in Table 6, which leverage crash performance dummies as instruments for Top Pick and only opposing vehicle weights for the weight differential. Due to computational cost, our estimation includes less granular fixed effects. Only model year dummies are included in Column 1. Column 2-3 additionally include vehicle type dummies and Column 3 introduces additional vehicle characteristics as controls. Average marginal effects (ME) are reported for each specification, which are all quantitatively similar to the linear estimates from Table 5.

The numerically calculated cross derivative between own-weight and opposing weights are also reported, labeled "Cross-Weight ME." This may be interpreted as the offsetting effect of the benefits of own-weight as opposing weight increases by 1,000 pounds. This interaction effect is a necessary component for the arms race feedback and, through nonlinearities of fatality risk, derives directly from the structure of the probit fatality risk model. These structural assumptions are necessary as direct estimation of interaction effects may be biased due to selection on own-weight. Our model allows us to pin down these efects under distributional assumptions leveraging only exogenous information on opposing vehicle weight. While acknowledging the endogeneity of own-weight in fatality outcomes, we compare these cross-derivatives to a directly estimated interaction effect in Appendix B. While the latter estimates may be biased, results are very comparable to the cross-effects reported in Table 6.

4.3 Combined Model

In a standard model of a vehicle arms race, consumers receive utility from weight for both the value placed on added capacity or cargo space and from the safety it provides (White, 2004; Li, 2012; Anderson and Auffhammer, 2014; Scott, 2022). However, these two mechanisms are difficult to disentangle from observational data on vehicle purchases. In this paper, we

form a thought experiment, whereby publicly available information on safety of a vehicle is adapted to account for vehicle weight, suggesting an additional lens in which preferences for weight as a safety attribute may be isolated from other factors.

Beginning with our linear specification of fatality risk, under the new ratings regime defined by Equation 4, we can write the counterfactual demand for vehicles as a function of vehicle weight and safety ratings as follows.

$$\begin{aligned} log(q_{jt}) =& \beta top^*_{jt} + \beta^w w_{jt} + \varepsilon_{jt} \\ =& \beta top_{jt} + \beta \frac{\gamma}{\delta}(w_{jt} - \bar{w}) + \beta^w w_{jt} + \varepsilon_{jt} \end{aligned} \tag{9}$$

This provides a framework for evaluating the extent to which preferences may be exogenously affected by an alternative ratings regime. Of interest is the additional demand for weight arising from the new methodology or, the combined parameter $\beta\gamma/\delta$. Assuming constant demand for weight due to other factors, β^w, we can compare these two weight parameters in order to measure the extent to which the size of vehicles might grow due to this change in ratings methodology.

Estimates from Table 4 and 5 project an added, implied preference for weight equal to $\hat{\beta}\hat{\gamma}/\hat{\delta} = (0.11)(0.00034)/0.00018 = 0.208$.[10] That is, these base estimates suggest that, under a re-calibrated ratings system that incorporates the role of weight, an additional 1,000 pound increase in vehicle weight would stimulate a 20 percent increase in demand for that vehicle. While we do not claim to identify the *direct* causal effect of weight on vehicle demand, estimates in Table 4 suggest a possible effect in the neighborhood of 70 percent per 1000 pounds. Comparing this to our estimate suggests an increase in preferences of nearly 30 percent following the counterfactual ratings redesign.

The counterfactual estimate we have just discussed calculates an additional role of vehicle weight in demand, through the safety ratings mechanism. However, as discussed in Section 3, a linear model of fatality risk does not allow for strategic complementarities in vehicle

[10]Note that Top Pick is estimated in 100s in Table 5, thus, the coefficient of interest here is $0.018/100 = 0.00018$.

weight preferences. If ratings properly accounted for relative vehicle size in car accidents, we would expect preferences for weight to be directly linked to average vehicle sizes, suggesting an interaction effect.[11] This is the main motivation for our probit fatality risk model. The counterfactuals of interest now directly derive from the following adapted demand equation.

$$\begin{aligned} log(q_{jt}) =& \beta top^*_{jt} + \beta^w w_{jt} + \varepsilon_{jt} \\ =& \beta top_{jt} + \beta \frac{\Phi\big(\alpha + \gamma(w_{jt} - \bar{w})\big) - \Phi\big(\alpha\big)}{\Phi\big(\alpha + \delta\big) - \Phi\big(\alpha\big)} + \beta^w w_{jt} + \varepsilon_{jt} \end{aligned} \tag{10}$$

Under the counterfactual rating design described in Equation 7, we analytically solve for both the implied added marginal effect of weight on demand, and the cross-derivative with average weight. That is, we derive the values: $\frac{\partial log(q_j)}{\partial top^*_j}\frac{\partial top^*_j}{\partial w_j}$ and $\frac{\partial}{\partial \bar{w}}\big(\frac{\partial log(q_j)}{\partial top^*_j}\frac{\partial top^*_j}{\partial w_j}\big)$. These estimates are reported in Table 7.

We analytically derive the solutions for each derivative and evaluate it at mean values of the covariates; own-weight and opposing weight are both evaluated at the mean weight in the accidents data of about 3,800 pounds.[12] Standard errors are derived by the delta method using the cluster-robust covariance matrices from Table 4 and 6, under independence between the demand and probit parameters. For the demand parameters, we only use estimate of specification 5 in Table 4 for each estimate, as demand estimates are relatively stable. Thus, differences in Columns 1-3 of Table 7 directly coincide with the probit specifications.

The estimates produced for the additional marginal effect of weight are quantitatively and qualitatively similar to those from the linear specification, suggesting a 20 to 23 percent increase in demand per 1,000 pound increase in own-vehicle size. Of particular interest is how these preferences for weight would vary when evaluated on different fleet weights. The second row in Table 7 suggests an additional 11-12 percent increase in preferences for weight per 1,000 pound increase in average fleet weight. For example, the marginal demand for a 1,000 pound increase in weight is 20 percent under an average fleet weight of 3,800 pounds; e.g., a

[11] The probit assumption is primarily used for convenience over alternative linear estimates on the weight interactions, due to the potential for selection on own-weight in fatality risk.

[12] I.e., marginal effects evaluated at $w_j - \bar{w} = 0$.

fleet of Ford Escapes. However, if fleet weight were to rise 500 pounds to 4,300—the approximate size of a Ford Explorer—marginal demand for weight will increase to about 25 percent.

Figure 2 depicts these effects visually. Panel A simply plots the fitted values from Equation 10 over a range of own-vehicle weights. The plotted curves, however, hold the effects of weight through direct preferences, β_w, constant.[13] The function is evaluated at two separate mean fleet weights, 3,000—roughly the size of a Toyota Corolla—and 5,000 pounds—about the size of a large (e.g., extended cab, 8-foot bed) Ford F150—while adjusting the intercepts to match mean demand for each curve. Estimates from Table 7 suggest an average slope for the demand curve under the 3,000 pound fleet equal to 13 percent, and for the 5,000 pound fleet equal to 35 percent.[14] Panel B plots the estimated marginal effects of weights directly, as a function of average fleet weight. While this plots the analytically derived marginal demand for weight directly, they align closely with our approximations from our average marginal effects described above. Furthermore, this graph clearly depicts how, under a weight-incorporated safety rating system, an arms race effect can spiral toward an inefficiently heavy fleet.

Finally, we translate these cross-effects on (logged) sales into the implied marginal effect of mean-weight on expected, purchased weight. We view this as a reduced-form interpretation of an arms race effect. Rather than imposing additional assumptions and estimating demand responses by simulation, we derive an expression for the implied effect directly. Specifically, we derive the following approximation.

$$\frac{\partial \mathbb{E}w_{j^*}}{\partial \bar{w}} \approx \mathbb{E}\Big[\frac{\partial}{\partial \bar{w}}\Big(\frac{\partial log(q_j)}{\partial top_j^*}\frac{\partial top_j^*}{\partial w_j}\Big)\Big] \cdot \text{Var}(w) \tag{11}$$

where $\mathbb{E}w_{j^*}$ denotes the expected value of the vehicle j^* which is chosen, the first term on the right-hand-side is the cross-effect of interest, and $\text{Var}(w)$ is the (sales-weighted) variance of vehicle weights in our sales data. The derivation for this expression is shown in Appendix

[13] We adjust the intercept term accordingly.

[14] Since the estimates in Table 7 condition on the average weight of 3,800 in the sample, we simply apply the relative weights of the 3,000 and 5,000 pound fleet weights to derive these effects (i.e, -800 and 1,200, or -0.8 and 1.2 in 1,000s).

C.

Estimates suggest that, under this revised, counterfactual safety rating system, which incorporates the objective role of relative weight into its ratings, consumers will respond to a 1,000 pound increase in mean fleet weight by increasing purchased weight by 70 pounds. Below these estimates are the derived multipliers, equal to $1/(1 - \mathbb{E}w_{j^*}/\partial\bar{w})$. This is the change in equilibrium weight that occurs when $\mathbb{E}w_{j^*}$ converges to $\bar{w}$ through the arms race feedback loop.[15] The results suggest that a manufactured arms race generated through weight-incorporated safety ratings will increase equilibrium fleet weight by 7-7.6 pounds. For perspective, at an average weight of 3,800 pounds in sample — approximately the size of a Toyota Rav4 or comparable mid-size SUV — the multiplier estimates translate into a nearly 300 pound increase in equilibrium weight. This is approximately the same as a shift from a fleet of Toyota Rav4 or similar mid-size SUVs to a fleet of Jeep Wranglers. Of course, this is just one channel in which an arms race could exist, while direct strategic responses to fleet weights may be even more pronounced.

4.4 Externalities

Table 7 reports a long-run equilibrium multiplier for vehicle weights implied from our model estimates. While this metric is useful for gathering intuition on the magnitude of the induced arms race effect, this value does not account for inertia in average fleet weight, generated due to slow turnover in the country's light duty vehicle fleet. When the supply-side offering of vehicles is held constant, in the long run, all vehicles in the fleet converge to a heavy, though static, mean weight. While this outcome is inefficient from an environmental or infrastructure perspective, under symmetric risks in traffic accidents, the equilibrium outcome on traffic fatalities effectively cancel out. That is, only relative weights matter.

In practice, manufacturers regularly upgrade their models, and turnover of the national fleet occurs at a relatively slow rate. This suggests a wedge between old and new vehicle weights might be permanently sustained under the ratings-induced arms race effect. We adapt the multiplier estimate to account for fleet turnover, and calculate this wedge in this section.

[15] Writing the arms race best response function as $\mathbb{E}w_{j^*} = a + b\bar{w}$, where $b \approx \partial\mathbb{E}w_{j^*}/\partial\bar{w}$, $a/(1-b)$ is the equilibrium weight and $1/(1-b)$ illustrates the equilibrium change in baseline weight, or the multiplier.

Let θ equal the fleet turnover rate. Specifically, we assume that each newly purchased vehicle replaces an older vehicle in the national fleet at random, and the fraction of vehicles replaced is defined by θ. Let w^* be the mean weight of new vehicles purchased and $\bar{w}_0$ be the preexisting mean weight of the fleet. When vehicles are scrapped and replaced independently of weight, the updated mean weights in any period are defined by:

$$\bar{w}_1 = \theta w^* + (1-\theta)\bar{w}_0$$

We acknowledge that the induced-arms race effect might lead to non-random scrappage of vehicles; e.g., smaller cars scrapped earlier. Therefore, we interpret the following calculations as general "ball-park" estimates, necessary in absence of a comprehensive model of scrappage. In a simple, linear arms race framework, we can write the expected vehicle weight of a new purchase as a function of the current fleet weight. That is, the best response function is:

$$\begin{aligned}\mathbb{E}w_{j^*} =& \bar{w}_0 + b\bar{w}_1 \\ =& \bar{w}_0 + b(\theta w^* + (1-\theta)\bar{w}_0)\end{aligned}$$

where the first equality illustrates that, under no feedback loop — i.e., when ratings do not incorporate weights such that $b = 0$ — expected weights are static (or evolve through factors other than the feedback loop). The second equality illustrates the manner in which new purchased weights, $\mathbb{E}w_{j^*}$, will converge to mean-weights for the replacement vehicles, w^*. Therefore, in equilibrium, expected new vehicle weights for any period are:

$$\mathbb{E}w_{j^*} = w^* = \bar{w}_0 \frac{1+(1-\theta)b}{1-\theta b}$$

This implies that the equilibrium wedge between new and existing vehicles in the fleet, following the ratings change, can be calculated as:

$$w^* - \bar{w}_1 = \bar{w}_0 \frac{(1-\theta)b}{1-\theta b} \tag{12}$$

Because this term is positive, it suggests that weight-incorporated safety ratings drive a per-

manent wedge between new vehicle purchase weight and the current fleet weight. One caveat to this argument is that it assumes that manufacturers also responds, unboundedly, to the feedback loop by continuously producing larger vehicles to satisfy demand. The supply-side response is not directly modeled, but constraints on current product offerings are implicit in our estimates through the restricted choice sets in our data.

The margin in which the externalities associated with the engineered arms race effect derives through its direct influence on new vehicle purchases. The counterfactual new vehicle is the source of the externality. Since the equality in Equation 12 is always positive, and larger vehicles impose higher risks to smaller vehicles, new vehicle purchases in this counterfactual scenario will generate a positive, average external cost on other drivers on the road. Importantly, this is a distributional safety externality: drivers of newer, heavier vehicles become safer while imposing the greatest risks on drivers of older, lighter vehicles. Vehicle weights in an arms race continously spiral upwards — bringing growing externalities such as environmental damage and road wear — while the individual who ultimately bears the risk is determined by turnover mechanics.

As an arms race only redistributes risks over time, we first focus on the immediate effect of the counterfactual new vehicles on drivers in the preexisting fleet. This calculation ignores other externalities associated with large vehicles and the long-run dynamics as the distribution of weights shifts over time.

$$\text{External Cost} = \#\text{New Purchases} \times \Delta Pr(\text{fatality} \mid w^* - \bar{w}) Pr(\text{accident}) \times VSL$$

For accident probability, we divide the total number of multi-vehicle accidents in the United States (NHTSA, 2022) by the total number of registered motor vehicles (DOT, 2022) and arrive at an estimate of 3.1 percent per year.[16] Anderson and Auffhammer (2014) make a similar calculation using a probability of multi-vehicle accident estimate of 3.65 percent derived

[16]NHTSA 2022, Table 33: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813656.pdf
DOT, 2022: https://www.fhwa.dot.gov/policyinformation/statistics/2022/pdf/mv1.pdf

from 2007 data. Our calculations treat turnover and accident risk as independent processes, though in practice a higher accident rate will induce additional scrappage. Crash-induced replacements should account for a portion of our turnover measure, and this relationship is held fixed in our calculation. More importantly, the extent to which weight in accidents affects scrappage rates is ignored as scrappage is not directly observable. Incorporating these effects would likely magnify the growth in fleet weights. Alternative turnover rates are explored below, though all independently of fleet weight.

Total number of new light-duty vehicle purchases come from the U.S. Bureau of Economic Analysis and was estimated at a total count of 15.8 million in 2024.[17] We apply the Environmental Protection Agency recommended value of statistical life of $10.7 million.[18] Finally, we apply the equilibrium vehicle weight wedge from Equation 12 using our Column 3 estimate from Table 7 for the feedback parameter and a base estimate for the turnover rate equal to 5.6 percent, and apply these estimates to our probit parameter values reported in Column 3 of Table 6 to derive the conditional probability of a traffic fatality.[19] The fatality risk estimate is calculated net of the baseline scenario of weightless safety ratings to obtain a marginal damage estimate, $\Delta Pr(\text{fatality} \mid \mathbb{E}w_{j^*} - \bar{w})$.

Our base estimates place the total external costs from the counterfactual safety ratings at $213.4 million per year. While this is a significant value, we acknowledge the imprecision of this estimate due to the assumptions on values outlined above, particularly with respect to a fixed turnover rate. It is likely that both fleet characteristics (e.g., weights) and supply-side information (e.g., safety ratings) influence this turnover. Firstly, new sales and old vehicle scrappage may not be one-to-one. Additionally, the rate in which the turnover may occur is likely stimulated by new information on safety. Due to data limitations, the degree to which safety ratings and fleet characteristics influence turnover as a whole is not directly modeled in this paper. To assess the sensitivity of turnover on these estimates (without fully endogenizing the prices) we gauge our primary estimate against alternative assumptions.

[17] https://fred.stlouisfed.org/series/ALTSALES?

[18] https://www.epa.gov/system/files/documents/2024-12/appendix-b-guidelines-for-preparing-economic-analyses_final_508-compliant.pdf

[19] Turnover is approximated as the number of new sales divided by the number of registered vehicles in the U.S.

Under an alternative estimate, the national scrappage rate is estimated at 4.5-4.6 percent according to S&P Global Mobility.[20] An estimate of $\theta = 0.045$ does not significantly affect our estimate, raising the external cost slightly to $216 million a year by partially increasing the weight wedge. However, this calculation lowers the turnover rate while holding the number of new purchases constant. Under one-to-one replacement, the revised number of new purchases should be 0.045 times the total fleet count. Applying this interpretation of turnover drops the implied new vehicle sales from 15.8 million to 12.7 million vehicles. This adjustment generate a significantly smaller external cost of $174 million.

In a more extreme setting, if we expect the new safety ratings to drastically increase turnover to a rate of 10 percent, external cost estimates increase to $363 million. The smaller wedge in vehicle weights in this setting is offset by a higher number of drivers wanting heavier vehicles. This is a non-monotonic relationship, however, as 100 percent turnover will ultimately lead to weight parity with an especially heavy vehicle fleet. While such an equilibrium might minimize traffic fatalities, it magnifies other externalities associated with large vehicles.

The new vehicles are the source of the external costs in a vehicle arms race, where the risk is shifted as the fleet evolves over time. Tracking new vehicle weights, w^*, over the dynamic growth in the overall fleet weight, $\bar{w}$, illustrates the redistribution of risks and the point at which the weight wedge for a single cohort of vehicles ultimately becomes negative. The equations above result in the following equilibrium process that describes the growth in average fleet weight.

$$\bar{w}_1 = \frac{1}{1 - \theta b} \bar{w}_0$$

In the context of this model, the switching point, where a vehicle that imposes net costs on other drivers transitions to one incurring those costs, occurs when its weight is less than the mean fleet weight — the weight wedge becomes negative. Given the counterfactual purchase weights from the safety ratings and the fleet weight process above, we estimate this switching

[20] https://www.spglobal.com/mobility/en/research-analysis/average-age-vehicles-united-states-2024.html?

point at about 17 years, at base parameter values. Over this horizon, cumulative costs from the new purchases will add up to about $1.8 billion, or $1.4 billion in net present value (at a 5% discount rate).

5 Conclusion

A vehicle arms race occurs when consumers rationally internalize vehicle size as a safety attribute, thus, purchasing heavier and heavier vehicles to mitigate their risks against other large cars on the road. Estimating this best response mechanism is difficult given the inherent feedback loops defining an arms race, though prior studies have provided evidence that such preferences likely exist in practice (e.g., Li, 2012, Scott, 2022). The social costs associated with this behavior has also been clearly documented in the literature (Crandall and Graham, 1989; Evans, 2001; Van Auken and Zellner, 2005; Anderson, 2008; Jacobsen, 2013; Anderson and Auffhammer, 2014), as heavier vehicles generally produce both environmental externalities and additional safety risks to other drivers.

When concerned about safety, a consumer may look to multiple sources of information to make a suitable vehicle choice. While they may directly internalize weight as a safety attribute, consumers looking to purchase a safe vehicle may also gather reputable information from published safety ratings. When these ratings contain objective information about the role of weight in vehicle accidents, the vehicle arms race can be exacerbated. Thus, a socially efficient testing approach may be one that excludes these effects.

This paper explores the efficiency improvements of the current safety ratings methodology which holds constant the role of weight in crash tests. We compare demand for vehicles under the current ratings system to a counterfactual rating which incorporates the marginal effects of relative weight on fatality risk. While the underlying function describing a safety rating may not be directly observed by the consumer, such a ratings system would create an increased, implied preference for weight by ranking heavier vehicles higher, holding other factors constant. Our estimates show that implied preferences for weight would increase by 30 percent following the change in methodology, increasing the marginal demand of a 1,000 pound increase in weight by 20 percent. Furthermore, we derive an indirect response to

fleet weights through the ratings mechanism, demonstrating a suggested 7 pound increase in pounds per 100 pound increase in average weight.

The findings in this paper illustrate the manner in which demand for weight may be exogenously influenced by the ratings agency. Although manipulation of preferences is not likely an objective, the controlled setting of the crash tests provide clear benefits by explicitly holding vehicle size constant. While standard fixed barrier tests only simulate collisions between equal sized vehicles, they are far less costly than an alternative which, for example, crashes vehicles of different sizes together. An important consideration is whether ratings agencies would have an incentive to adopt weight-incorporated ratings should testing become more affordable. High demand by consumers and insurance companies for an accurate depiction of risk could easily drive testing methodology in that direction; a practice, that this paper illustrates, could produce costly outcomes, by exacerbating the effect of a vehicle arms race.

References

Alberini, Anna and David H. Austin. 1999. "Economic Valuation of Automobile Injury Reduction: Estimates from Automobile Demand Models." *Journal of Risk and Uncertainty* 18 (1):25–48.

Allcott, Hunt. 2013. "The Welfare Effects of Misperceived Product Costs: Data and Calibrations from the Automobile Market." *American Economic Journal: Economic Policy* 5 (3):30–66.

Allcott, Hunt and Nathan Wozny. 2014. "Gasoline Prices, Fuel Economy, and the Energy Paradox." *Review of Economics and Statistics* 96 (5):779–795.

Anderson, Michael. 2008. "Safety for Whom? The Effects of Light Trucks on Traffic Fatalities." *Journal of Health Economics* 27 (4):973–989.

Anderson, Michael L. and Maximillian Auffhammer. 2014. "Pounds That Kill: The External Costs of Vehicle Weight." *Review of Economic Studies* 82 (2):535–571.

Arato, Hiroki and Tomoya Nakamura. 2022. "Welfare Effects of Partial Publicity and Partial Transparency in Endogenous Information Acquisition." *Available at SSRN 4155967* .

Busse, Meghan, Christopher R. Knittel, and Florian Zettelmeyer. 2013. "Are Consumers Myopic? Evidence from New and Used Car Purchases." *American Economic Review* 103 (1):220–256.

Cohen, Alma and Rajeev Dehejia. 2004. "The effect of automobile insurance and accident liability laws on traffic fatalities." *The journal of law and economics* 47 (2):357–393.

Crandall, Robert W. and John D. Graham. 1989. "The Effects of Fuel Economy Standards on Automobile Safety." *Journal of Law and Economics* 32:97–118.

Evans, Leonard. 2001. "Causal Influence of Car Mass and Size on Driver Fatality Risk." *American Journal of Public Health* 91 (7):1076–1081.

Greenstein, Shane and Feng Huang. 2017. "Vehicle Safety and Manufacturer Reputation: Evidence from Vehicle Recalls." *Journal of Industrial Economics* 65 (4):749–786.

Hall, Jonathan D. and Joshua Madsen. 2022. "Can Behavioral Interventions Be Too Salient? Evidence from Traffic Safety Messages." *Science* 376 (6591).

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

Hoekstra, Mark, Steven L. Puller, and Jeremy West. 2017. "Cash for Corollas: When Stimulus Reduces Pollution." *American Economic Journal: Applied Economics* 9 (3):1–35.

Houde, Sébastien. 2014. "How consumers respond to environmental certification and the value of energy information." Tech. rep., National Bureau of Economic Research Cambridge, MA, USA.

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

Ippolito, Pauline M and Alan D Mathios. 1990. "Information, advertising, and health choices: a study of the cereal market." *Rand J. Econ* 21:211–246.

Jacobsen, Mark R. 2013. "Fuel Economy and Safety: The Influences of Vehicle Class and Driver Behavior." *American Economic Journal: Applied Economics* 5 (3):1–26.

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

Kamenica, Emir. 2019. "Bayesian persuasion and information design." *Annual Review of Economics* 11 (1):249–272.

Landau, L.D. and E.M. Lifshitz. 1976. *Mechanics*, *Course of Theoretical Physics*, vol. 1. Oxford: Pergamon Press, 3rd ed.

Li, Shanjun. 2012. "Traffic Safety and Vehicle Choice: Quantifying the Effects of the 'Arms Race' on American Roads." *Journal of Applied Econometrics* 27:34–62.

Luca, Michael. 2011. "Reviews, reputation, and revenue: The case of Yelp. com (No. 12-016)." *Harvard Business School* .

Morris, Stephen and Hyun Song Shin. 2002. "Social Value of Public Information." *American Economic Review* 92 (5):1521–1534. URL https://www.aeaweb.org/articles?id=10.1257/000282802762024610.

Peltzman, Sam. 1975. "The Effects of Automobile Safety Regulation." *Journal of Political Economy* 83:677–726.

Peterson, Steven, George Hoffer, and Edward 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.

Sallee, James M. 2014. "Rational Inattention and Energy Efficiency." *Journal of Law and Economics* 57 (3):781–820.

Sallee, James M., Sarah E. West, and Weiwei Fan. 2016. "Do Consumers Recognize the Value of Fuel Economy? Evidence from Used Car Prices and Gasoline Price Fluctuations." *Journal of Public Economics* 135:61–73.

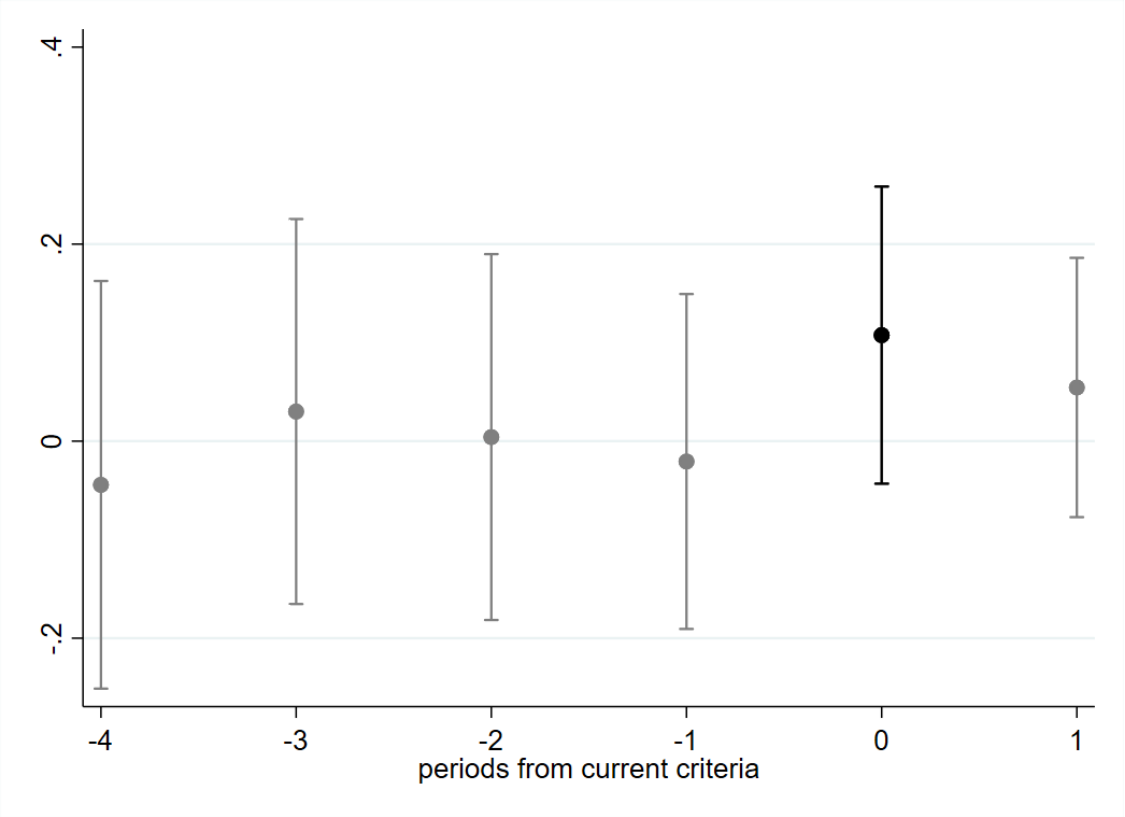
Scott, Jonathan B. 2022. "Pounds That Save: The Role of Preferences for Safety in Demand for Large Vehicles." *Journal of Law and Economics* 65 (3):555–579.

Van Auken, M. and J. W Zellner. 2005. "An Assessment of the Effects of Vehicle Weight and Size on Fatality Risk in 1985 to 1998 Model Year Passenger Cars and 1985 to 1997 Model Year LIght Trucks and Vans." *SAE International* .

White, Michelle J. 2004. "The 'Arms Race' on American Roads: the Effect of SUVs and Pickup Trucks on Traffic Safety." *Journal of Law and Economics* 47:333–356.

Figures

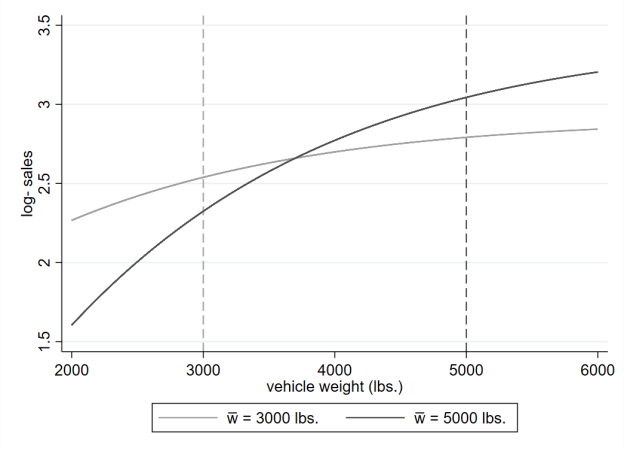
Figure 1: Placebo Lag/Leading Criteria Changes



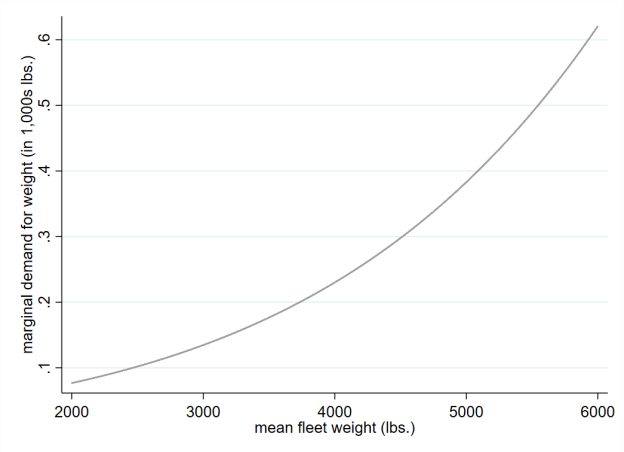
Note: This figure plots the coefficients of several placebo top safety pick ratings that assign based on previous or future years' criteria. The coefficient at period zero defines assignment based on the true and current criteria for a given model year.

Figure 2: Effect of Mean-Weight on Demand From Ratings Change

(a) Vehicle Demand by Own-Weight



(b) Marginal Demand for Weight by Mean Fleet Weight



Note: Panel a presents the fitted values of a vehicle demand model, conditional on vehicle weight, through the safety ratings channel. Panel b presents the marginal effects of weight on demand, conditional on mean fleet weight.

Tables

Table 1: IIHS Top Pick Criteria

Model Year:	2015	2016	2017	2018	2019
Test 1: Driver Side Small Overlap Front	Acceptable	Good	Good	Good	Good
Test 2: Passenger Side Small Overlap Front	N/A	N/A	N/A	N/A	Acceptable
Test 3: Moderate Overlap Front	Good	Good	Good	Good	Good
Test 4: Original Side	Good	Good	Good	Good	Good
Test 5: Roof Strength	Good	Good	Good	Good	Good
Test 6: Head and Seat Restraint Test	Good	Good	Good	Good	Good
Test 7: Headlight Rating	N/A	N/A	N/A	Acceptable	Acceptable
Test 8: Front Crash Prevention Technology	N/A	Basic	Advanced	Advanced	Advanced

Note: Above is the minimum rating requirement in each model year for the 8 primary crash tests to achieve IIHS top pick status. Tests 1-7 are evaluated on a scale: Poor, Marginal, Acceptable, Good. Column 8 is evaluated on a scale: Not Equipped, Basic, Advanced, Superior. "N/A" indicates that there is no minimum requirement for the test.

Table 2: Summary of Key Variables

Variable	Top	Non–Top	Full sample
Curb weight (lbs)	3,659.13	3,980.14	3,858.27
	(556.82)	(897.83)	(801.27)
MSRP ($)	33,787.57	36,344.31	35,374.51
	(13,015.25)	(15,087.95)	(14,390.46)
Miles Per Gallon (MPG)	29.51	24.07	26.13
	(13.49)	(9.86)	(11.67)
Horsepower	228.18	257.65	246.47
	(68.41)	(88.53)	(82.73)
Cumulative sales	2,395,980	2,858,694	5,254,674
Number of crashes	354,407	392,157	746,564
Total deaths	426	516	942

Note: The table reports average levels of key vehicle characteristics and their standard deviations, where an observation is a unique vehicle model. Variables are additionally tabulated by their IIHS award status. Total sales, total number of crashes, and total number of traffic fatalities are reported for each category.

Table 3: Highest Expected Ratings Flips

Top-Pick	Make-Model	Curb Weight (lbs)	Relative to Mean Weight
MY: 2015			
Yes	Chevrolet Spark	2400	-33%
Yes	Honda Fit	2600	-28%
Yes	Hyundai Elantra	2800	-22%
Yes	Kia Soul	2800	-22%
Yes	Subaru BRZ	2800	-22%
No	Ford Expedition EL	5900	64%
No	GMC Yukon XL	5700	58%
No	Ford Expedition	5600	56%
No	GMC Yukon	5500	53%
No	Chevrolet Tahoe	5300	47%
MY: 2016			
Yes	Scion IA	2400	-35%
Yes	Chevrolet Sonic	2800	-24%
Yes	Honda Civic	2800	-24%
Yes	Kia Soul	2800	-24%
Yes	Mazda CX3	2900	-22%
No	Lincoln Navigator L	6100	65%
No	Lincoln Navigator	5900	59%
No	Cadillac Escalade ESV	5900	59%
No	Ford Expedition EL	5900	59%
No	GMC Yukon XL	5700	54%
MY: 2017			
Yes	Toyota Yaris IA	2400	-35%
Yes	Honda Civic	2800	-24%
Yes	Hyundai Elantra	2800	-24%
Yes	Mazda CX3	2800	-24%
Yes	Hyundai Elantra GT	2900	-22%
No	Nissan Titan XD	6300	70%
No	Lincoln Navigator L	6200	68%
No	Cadillac Escalade ESV	5900	59%
No	Ford Expedition EL	5900	59%
No	Lincoln Navigator	5900	59%
MY: 2018			
Yes	Nissan Kicks	2700	-27%
Yes	Kia Rio	2700	-27%
Yes	Kia Forte	2800	-24%
Yes	Hyundai Elantra	2800	-24%
Yes	Toyota Corolla	2900	-22%
No	Cadillac Escalade ESV	6000	62%
No	GMC Yukon XL	5700	54%
No	Chevrolet Suburban	5700	54%
No	Cadillac Escalade	5700	54%
No	Nissan Titan	5600	51%
MY: 2019			
Yes	Hyundai Accent	2700	-29%
Yes	Kia Rio	2700	-29%
Yes	Nissan Kicks	2700	-29%
Yes	Hyundai Veloster	2800	-26%
Yes	Kia Forte	2900	-24%
No	GMC Yukon XL	5800	53%
No	Cadillac Escalade	5700	50%
No	Chevrolet Suburban	5700	50%
No	Nissan Titan	5600	47%
No	GMC Yukon	5500	45%

Table 4: Vehicle Demand

	(1)	(2)	(3)	(4)	(5)	(6)
Top Pick	0.158***	0.105**	0.147**	0.115*	0.111**	0.110**
	(0.0437)	(0.0513)	(0.0617)	(0.0599)	(0.0559)	(0.0553)
Lagged Top Pick		0.0820	0.0780	-0.0104		
		(0.0517)	(0.0606)	(0.0638)		
MSRP ($10,000s)						-0.146*
						(0.0858)
Curb Weight (1,000 lbs)						0.719***
						(0.251)
MPG						0.0149
						(0.0118)
Horsepower (100s)						-0.172
						(0.203)
City FE	Yes	Yes	-	-	-	-
Make × Year FE	Yes	Yes	-	-	-	-
Make × Model FE	Yes	Yes	Yes	Yes	Yes	Yes
City × Make × Year FE	-	-	Yes	Yes	Yes	Yes
Test Performance Dummies	-	-	-	Yes	Yes	Yes
Observations	32581	32581	32581	32581	32581	32581

Note: Standard errors clustered at model-year level. Outcome variable, log(share), is logged aggregated quantites of model j in model year t, purchased in city c. This table shows the estimated effect of IIHS top pick on vehicle demand across different model specifications.

Table 5: Linear Fatality Risk Model

	(1)	(2)	(3)	(4)	(5)
Top Pick/100	-0.01133	-0.01277*	-0.01833**	-0.01833**	-0.01844**
	(0.00726)	(0.00742)	(0.00866)	(0.00890)	(0.00897)
Opposing Vehicle Weight (1,000 lbs)		0.00034***	0.00034***	0.00034***	0.00034***
		(0.00004)	(0.00004)	(0.00004)	(0.00004)
Curb Weight (1,000 lbs)		-0.00027***	-0.00031***	-0.00027***	-0.00027***
		(0.00010)	(0.00007)	(0.00010)	(0.00010)
MSRP ($10,000s)		-0.00003		-0.00003	-0.00003
		(0.00005)		(0.00005)	(0.00005)
MPG		-0.00000		-0.00000	-0.00000
		(0.00001)		(0.00001)	(0.00001)
Speed Limit		0.00004***		0.00004***	0.00004***
		(0.00000)		(0.00000)	(0.00000)
Vehicle-type FE	Yes	Yes	Yes	Yes	-
Model Year FE	Yes	Yes	Yes	Yes	-
Vehicle-type × Model-year FE	-	-	-	-	Yes
Crash Year × County FE	Yes	Yes	Yes	Yes	Yes
Top Pick	OLS	OLS	2SLS	2SLS	2SLS
Observations	544340	544340	544340	544340	544340

Note: Standard errors clustered at crash level. Outcome variable, Fatality, is a binary variable that equals 0 for no death count and 1 for positive number of death counts. This table shows the estimated effect of top pick on fatality risk across different model specifications. Crash test outcomes are used as instruments for top safety pick in Columns 3-5.

Table 6: Probit Fatality Risk Model

	(1)	(2)	(3)
Weight Differential (1,000 lbs)	-0.16241***	-0.16168***	-0.15073***
	(0.01192)	(0.01280)	(0.01329)
Top Pick/100	-10.12770***	-9.25490**	-8.36151**
	(3.62114)	(3.66504)	(4.02578)
MSRP ($10,000s)			-0.03330
			(0.02287)
MPG			-0.00444
			(0.00426)
Speed Limit			0.02672***
			(0.00147)
Weight ME	-.000375	-.000373	-.0003235
Top ME	-.023383	-.021349	-.017945
Cross-Weight ME	-.0001862	-.0001818	-.0001636
Vehicle-type FE		Yes	Yes
Model Year FE	Yes	Yes	Yes
Observations	544408	544408	544408

Note: Standard errors clustered at crash level. The results are from IV probit estimation where outcome variable, Fatality, is a binary variable that equals 0 for no death count and 1 for positive number of death counts. The table reports the estimated effect of top pick and weight differences on fatality risk across different model specifications. The crash test outcomes are used as instruments for top safety pick and only opposing vehicle weight is leveraged for the weight differences.

Table 7: Counterfactual Calculation of Added Value of Weight

	(1)	(2)	(3)
$\mathbb{E}\big[\frac{\partial log(q_j)}{\partial top_j^*} \cdot \frac{\partial top_j^*}{\partial w_j}\big]$	0.207	0.223	0.228
	(0.121)	(0.136)	(0.150)
$\mathbb{E}\big[\frac{\partial}{\partial \bar{w}}\big(\frac{\partial log(q_j)}{\partial top_j^*} \cdot \frac{\partial top_j^*}{\partial w_j}\big)\big]$	0.109	0.117	0.119
	(0.065)	(0.074)	(0.081)
$\frac{\partial \mathbb{E} w_{j*}}{\partial \bar{w}}$	0.065	0.070	0.071
multiplier	1.070	1.075	1.076
Demand Model:			
Test Performance Dummies	Yes	Yes	Yes
Make × Model FE	Yes	Yes	Yes
City × Make × Year FE	Yes	Yes	Yes
Crash Fatality Model:			
Model Year FE	Yes	Yes	Yes
Vehicle-type FE		Yes	Yes
Controls			Yes

Note: Counterfactual estimates analytically calculate the added preference for weight through the adjusted top safety pick mechanism. The results are derived from the demand estimates in Table 4 and fatality risk estimates in Table 6. Standard errors are approximated by the delta method using the cluster robust covariance matrices estimated from the two models.

Appendices

A NHTSA Ratings

The analysis in the main text estimates a response to published safety information using IIHS safety ratings due to their completeness in testing across models. IIHS awards top safety picks every year based on their comprehensive testing procedure. In contrast, for NHTSA, a selective sample of models are chosen each year, with priority placed on vehicles with high expected sales, and those whose models have been significantly altered for that model year.[21] A total of 37 models were chosen for NHTSA safety testing for model year 2025.[22] The implication is that consumers are usually guided to lagged ratings for a given model when a current rating is not available. Which previous year's rating is reported for a particular model and vintage is not directly observable from the NHTSA API, as we are only able to observe new tests and new ratings.

For consistency, we attempt to replicate our findings on NHTSA data by assuming that models may retain their old ratings for up to three model years. This assumption is only imposed when a new rating for a given model year is not reported in the data and one within the previous 3 model years is observed. We consider this to be a conservative assumption, however, it will still likely introduce additional measurement error into the estimates. The data are fuzzy matched to our sales data using vehicle name descriptions. We expect this to additionally introduce some measurement error, though likely random and, thus, should attenuate our estimates.

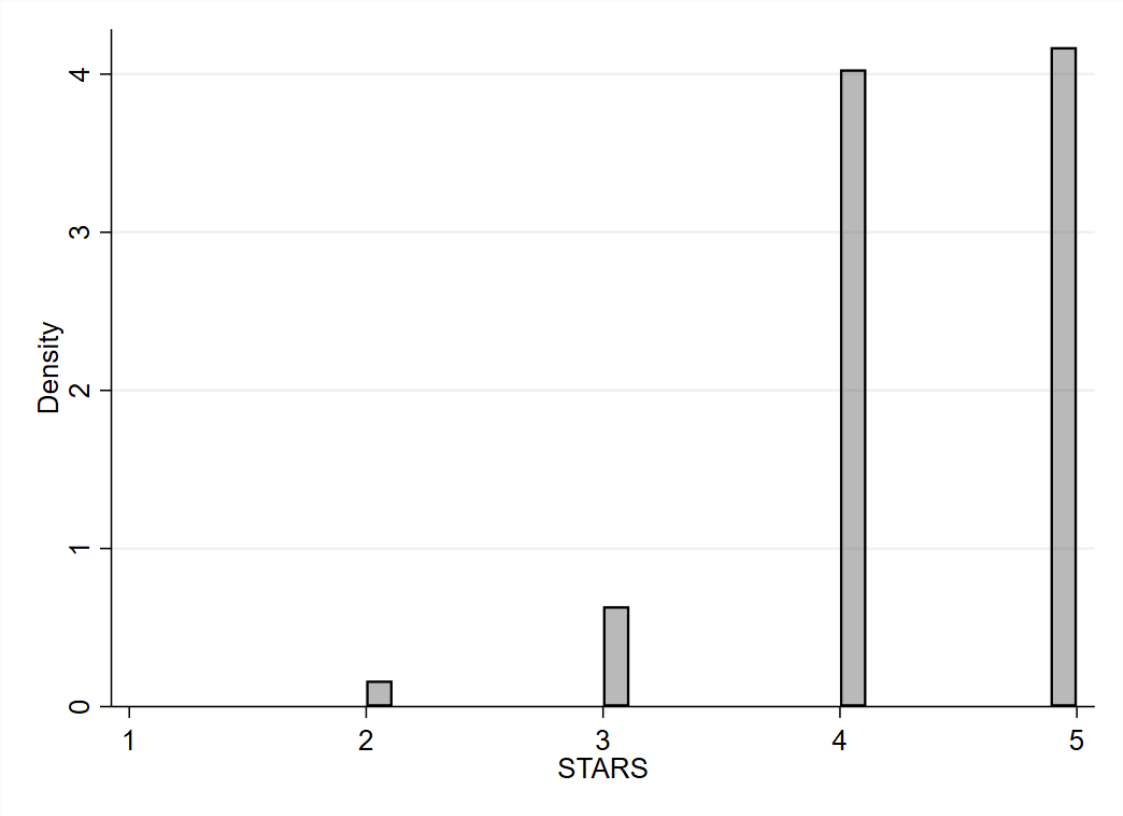
The distribution of NHTSA 5-star safety ratings is displayed in Figure A.1, where the unit of observation is a model-by-year. The vast majority of ratings are either 4 or 5 stars, and this

[21] https://www.caranddriver.com/features/g35634275/what-to-know-about-the-wrecks-behind-the-ratings-feature

[22] https://www.nhtsa.gov/press-releases/nhtsa-2025-vehicles-5-star-safety-ratings-testing

is plausibly due to the selection process of models chosen to be tested. The cutoffs at each star is determined based on a model's Vehicle Safety Score (VSS), a value which combines the results of various tests.

Figure A.1: Distribution of NHTSA Safety Ratings



Note:

To form a specification analogous to those in the main text, we focus on the effect of a 5-star award; we view this as analogous to a top safety pick. We control for the VSS running variable and estimate a similar fixed effects specification.[23] The results are reported in Table A.1. While our estimates are not statistically significant in any of the specifications, they still appear economically meaningful. The results of the main two-way fixed effect suggest 8-9 percent increase in sales for 5-star rated vehicles. Though the information treatment differs, this effect is similar in magnitude to the estimates uncovered for the IIHS top safety

[23]The data constuction process described above likely contributes to significant measurement error in the VSS values and the overall safety ratings. A pure regression discontinuity design does not provide robust results. Therefore, we focus on the fixed effects specification used in the paper.

picks.

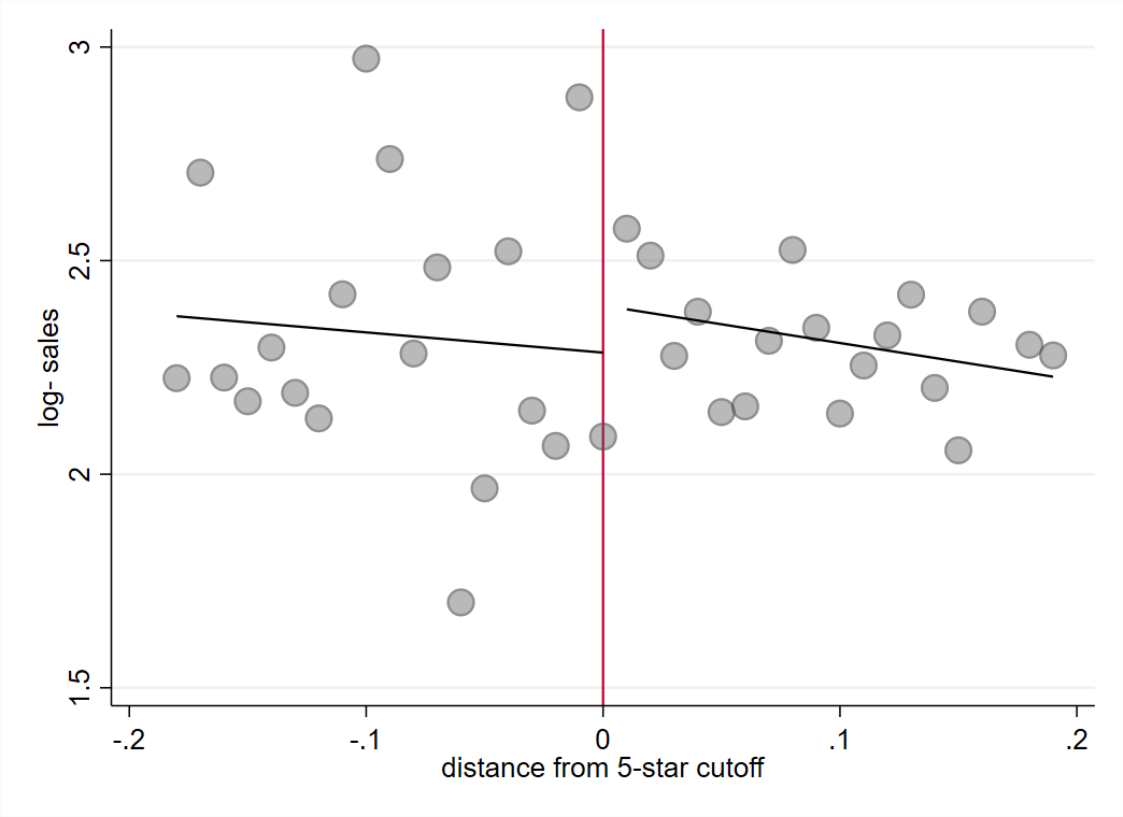
Table A.1: Vehicle Demand

	(1)	(2)	(3)
5-Stars	0.120**	0.135**	0.130*
	(0.0550)	(0.0672)	(0.0665)
VSS	-0.114	-0.0760	-0.0737
	(0.232)	(0.292)	(0.289)
City FE	Yes	-	-
Make × Year FE	Yes	Yes	Yes
Make × Model FE	Yes	-	-
City × Make × Year FE	-	Yes	Yes
Vehicle Characteristics	-	-	Yes
Observations	24949	24949	24949

Note: Standard errors clustered at model-year level. The outcome variable is logged aggregated quantites of model j in model year t, purchased in city c. This table shows the estimated effect of a NHTSA 5-star rating on vehicle demand across different model specifications.

The fixed effects estimator applied for the model in Table A.1 for consistency with our empirical approach in the main text. The estimates on 5-Stars are interpreted relative to all other vehicles in the sample, of varying ratings. An alternative is to conduct a regression discontinuity design on 4 versus 5 star vehicles. This approach produces statistically insignificant results — likely due to the degree of measurement error — with a coefficient estimates between 12 and 14 percent, depending on specification. Figure A.2 provides some evidence of a discontinuity, where positive distances from the VSS threshold depict sales of 5-star vehicles.

Figure A.2: Distribution of NHTSA Safety Ratings



Note: The above figure plots sales over the distance from the VSS threshold, determining 5 versus 4-star ratings. A 5-star vehicle is defined with a value of VSS equal to 0.675 or lower (distances are defined by 0.675-VSS).

B Interaction Effects in Linear Model

In Table 6 of the main text, we report average cross derivative estimates between the two weights involved in the two-vehicle accidents. This effect relies on exogenous variation in opposing vehicle weight, but also the structure of the probit model. Under a hypothetical, exogenous own-vehicle weight, we would be able to credibly identify an interaction effect directly; however, in practice, it is very plausible that selection plays a role in vehicle accidents, making endogeneity of own-weight an issue. While marginal effects of each vehicles' weight reported in Table 5 are similar in magnitude, our strategy opts to rely on nonlinearities of the probit model rather than introducing potential endogeneity into our estimates.

As a check, we provide a version of our linear model that estimates an interaction effect between the two vehicle weights directly. Under exogenous weights, we would expect this

interaction effect to be similar to the cross-derivative reported in Table 6. The results are presented in Table A.2.

Table A.2: Linear Fatality Risk Model w/ Weight Interactions

	(1)	(2)	(3)
Top Pick/100	-0.01832**	-0.01832**	-0.01842**
	(0.00866)	(0.00890)	(0.00897)
Weight × Opp. Weight	-0.00011**	-0.00011**	-0.00011**
	(0.00005)	(0.00005)	(0.00005)
Opposing Vehicle Weight (1,000 lbs)	0.00074***	0.00073***	0.00074***
	(0.00021)	(0.00021)	(0.00021)
Curb Weight (1,000 lbs)	0.00011	0.00014	0.00014
	(0.00020)	(0.00022)	(0.00022)
MSRP ($10,000s)		-0.00003	-0.00003
		(0.00005)	(0.00005)
MPG		-0.00000	-0.00000
		(0.00001)	(0.00001)
Speed Limit		0.00004***	0.00004***
		(0.00000)	(0.00000)
Vehicle-type FE	Yes	Yes	-
Model Year FE	Yes	Yes	-
Vehicle-type × Model-year FE	-	-	Yes
Crash Year × County FE	Yes	Yes	Yes
Top Pick	2SLS	2SLS	2SLS
Observations	544340	544340	544340

Note: Standard errors clustered at crash level. Outcome variable, Fatality, is a binary variable that equals 0 for no death count and 1 for positive number of death counts.

Estimates in Table A.2 replicate the specifications in Columns 3-5 of Table 5. The estimated interaction effects are similar and not statistically distinguishable from the numerically calculated cross-derivatives in Table 6. Given this information, we do not expect the probit structure — leveraged for interaction effects in this paper — to significantly alter our main results.

C Derivation of Marginal Effect of Average Weights

Our objective is to form a simple, point estimate for the marginal effect of average weight on purchased vehicle weight without significant structural assumptions. We provide a back-of-the-envelope estimate for this effect in Section 4. This appendix provides the analytical

derivation.

First, make note that our demand equation of interest is estimated on logged sales. This simplifies to a standard logit model, as the denominator of market shares are constant within market, and absorbed by our fixed effects. Denote cumulative market-level vehicle sales as $Q = \sum_{j \in J} q_j$, equal to the sum of all vehicle sales in that market. Thus, the market share for model j is q_j/Q. Denote the expected value of the weight of chosen vehicle j^* as:

$$\mathbb{E}[w_{j^*}] = \sum_{j \in J} \frac{w_j q_j}{Q}$$

Of interest is the marginal effect of average weight, $\bar{w}$, on expected chosen weight, as measured through the, counterfactual, top safety pick mechanism.

$$\frac{\partial \mathbb{E}[w_{j^*}]}{\partial \bar{w}} = \frac{\sum_{j \in J} \frac{\partial q_j}{\partial \bar{w}} w_j}{Q} - \frac{\sum_{j \in J} q_j w_j}{Q} \cdot \frac{\sum_{j \in J} \frac{\partial q_j}{\partial \bar{w}}}{Q}$$

Counterfactual demand under the new safety rating is defined by Equation 9 in the main text, where $\bar{w}$ influences preferences through top_j^*. As quantities are expressed in logs in our regressions, marginal effects on levels can be expressed as:

$$\frac{\partial q_j}{\partial \bar{w}} = q_j \frac{\partial log(q_j)}{\partial \bar{w}} \quad or \quad = \beta q_j \frac{\partial top_j^*}{\partial \bar{w}}$$

Our approximation comes from the first term on the right-hand-side, as these involve the average marginal effects derived in the paper. Replacing these terms we now have:

$$\frac{\partial \mathbb{E}[w_{j^*}]}{\partial \bar{w}} = \frac{\sum_{j \in J} q_j \frac{\partial log(q_j)}{\partial \bar{w}} w_j}{Q} - \frac{\sum_{j \in J} q_j w_j}{Q} \cdot \frac{\sum_{j \in J} q_j \frac{\partial log(q_j)}{\partial \bar{w}}}{Q}$$

Table 7 reports both the marginal effect of weight on logged quantities, and the cross-derivative with average weights. A linear approximation of logged-quantities on the interaction between weights can be expressed as a constant coefficient on the interaction term

$w_j \times \bar{w}$; for example, in the reduced form expression $log(q_j) = \gamma_0 + \gamma_1 w_j + \gamma_2 w_j \times \bar{w} + \nu_j$, where baseline effects of $\bar{w}$ are not relevant in the relative demand for vehicle j. The estimate of interest in Table 7 is $\gamma_2 \approx \frac{\partial^2 log(q_j)}{\partial w_j \bar{w}}$. Under this approximation, let $\frac{\partial log(q_j)}{\partial \bar{w}} = \gamma_2 w_j$, then we have:

$$\begin{aligned}
\frac{\partial \mathbb{E}[w_{j^*}]}{\partial \bar{w}} &= \gamma_2 \frac{\sum_{j \in J} q_j w_j^2}{Q} - \gamma_2 \frac{\sum_{j \in J} q_j w_j}{Q} \cdot \frac{\sum_{j \in J} q_j w_j}{Q} \\
&= \gamma_2 \cdot (\mathbb{E}[w_j^2] - \mathbb{E}[w_j]^2) \\
&= \gamma_2 \cdot \mathrm{Var}(w)
\end{aligned}$$

Following this formulation, we calculate the variances of purchased vehicle weights from our data using observed market shares and incorporate our average cross-derivative estimates from Table 7 to estimate the implied marginal effect of average weight on chosen vehicle weight.