Essays on the Role of Metrics in Innovation

by

Jane Yajie Wu

B.Com., Queen's University (2012)
B.A., Queen's University (2012)
S.M., Massachusetts Institute of Technology (2018)

Submitted to the Department of Management
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Management

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2022

Author . . . .
Department of Management
April 8, 2022

Certified by . . . .
Scott Stern
David Sarnoff Professor of Management of Technology
Thesis Supervisor

Accepted by . . . .
Catherine Tucker
Sloan Distinguished Professor of Management
Professor, Marketing
Faculty Chair, MIT Sloan PhD Program

Essays on the Role of Metrics in Innovation

by

Jane Yajie Wu

Submitted to the Department of Management
on April 8, 2022, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Management

Abstract

This dissertation consists of three essays studying the role of metrics in the process of innovation. Scientific and technical metrics are trusted as objective and consistent arbiters of knowledge, and as a result, are typically taken as given without much question. Yet at the same time, these metrics are chosen at a given point in time under imperfect information. The motivation of this work is to understand how such metrics influence the ideas production process, and ultimately, who benefits from innovative effort. In the first essay, I define and delineate the role of metrics in innovation from other forms of quantification in organizations. I synthesize prior work to develop a typology of mechanisms that metrics can involve, highlighting how metrics are used at different junctures in the innovation process. The second essay explores the impact of introducing a new metric on the rate and direction of innovation. I study the setting of US automotive safety, finding that the introduction of the side impact dummy as a metric reduced overall fatalities but also led to disproportionate benefits for occupants similar to the metric itself. Moreover, firms responded heterogeneously, suggesting that metrics can profoundly affect the innovation trajectories of firms. In the third essay, I analyze whether it is possible to move firms away from a metric that has become a key focusing device for R&D within an industry. I use a policy shock to estimate the effects of the "removal" of watts as a metric within the domestic vacuum cleaner industry. I find that rather than investing in new metrics, firms reduced their R&D in the focal area and shifted efforts to adjacent, unregulated product areas.

Thesis Supervisor: Scott Stern
Title: David Sarnoff Professor of Management of Technology

I dedicate this work to the memory of my grandfathers, Ma Wengui and Wu Zhaojiong, who believed in the value of scholarly pursuits and planted the seed for this endeavor years ago.

Acknowledgments

There is simply no metric to capture the impact that my advisors —Scott Stern, Pierre Azoulay, Fiona Murray and Timothy Simcoe —have had on me. I am grateful for their generous time, mentorship and invaluable insights.

Through his contagious energy, creativity and dedication to pursuing big questions, Scott has had a tremendous impact on my growth, both professionally and personally. This dissertation would not have been possible without his conviction that I could push further, and the inspiration from our countless conversations over the years. Learning what constitutes good research from Pierre has been a real privilege. I am grateful for his patient encouragement, for stressing the importance of understanding the context, and for sharing his passion for the process of scientific discovery. Fiona has provided so much wisdom not only in terms of asking sharp, clarifying questions on my research, but also on balancing professional ambitions with personal ones. I am grateful to her for being a generous champion, a trailblazer, and a role model. Finally, I have benefited enormously from my interactions with Tim (not to mention his indispensable econometrics class!). His actionable advice and ability to surface non-obvious connections has improved my research, and his own work on standards has influenced my thinking on metrics. Thank you as well to Cathy, Andi and Steph for giving my family and I a feeling of home in Boston throughout the years.

My non-linear path to the PhD began at the Next 36. There, I met Ajay Agrawal and Joshua Gans who gave me my first taste of the economics of innovation and entrepreneurship. I want to thank them for their kind support throughout the zigs and zags of my entrepreneurial and academic pursuits. There are a number of other scholars I am indebted to for their feedback and friendship, but I wanted to espe-

cially thank Abhishek Nagaraj, Ariel Stern, Avi Goldfarb, Catherine Tucker, Dan Fehder, Danielle Li, Ed Roberts, Eric von Hippel, Ezra Zuckerman Sivan, Florenta Teodoridis, Hong Luo, Jeff Furman, Jorge Guzman, Josh Krieger, Laurina Zhang, Mercedes Delgado, Pian Shu and Phil Budden. Special thanks to all of my TIES PhD colleagues for challenging my ideas and for the many laughs over the years. Thank you to Hillary Ross and Davin Schnappauf for creating a place of support for us PhD students. Thank you to Natalia Kalas, Kim McGrath, and Judy Graham-Robey for their kind help in navigating Sloan.

My deepest thanks go to my parents for their enormous sacrifices and unwavering support. Everything I have been able to do is thanks to what they have given to me. To Annie, I am lucky to have such an awesome sister. Part of the fun of Boston was being close enough to go on adventures with you. To my in-laws, thank you for your support, for crossing oceans to help us, and for reminding me to see the bigger picture. Finally, thank you to Christian who taught me about cryptographic hash functions: 3ef53670aa75a7ec1138a764c3ff3dac9a7c4a4d09b1a628407e1493d0ac17a9

Contents

List of Figures

List of Tables

Chapter 1

Introduction

From the Ancient Egyptian cubit to the Académie des Sciences mètre to the quantum qubit, metrics seem like a common thread throughout scientific and technological progress. Metrics can facilitate the diffusion, comparison, and recombination of ideas, across greater distances in geographic and domain space than otherwise would be possible. The use of metrics allows innovation to be centered on experiments rather than experience, expanding who can participate in the process. Put differently, metrics have become a prerequisite for understanding in modern scientific and technological innovation. Yet despite this level of importance, metrics have not always been endemic to ideas production. In fact, for many periods throughout human history, knowledge relied on qualitative reasoning rather than numeric measures.[1] What this suggests is that although scientific and technical metrics have unlocked significant progress in innovation, the entwinement of metrics with knowl-

[1] Scholars such as Aristotle and Plato for instance, reasoned that while the human mind was highly capable, the five senses were simply too mortal to accurately gauge the heterogeneous and evolving qualities of Nature. Instead during this period, knowledge was formed through qualitative observations and inductive reasoning, with a skepticism towards quantitative measures (Crosby, 1997).

edge is a social construct that has become the norm relatively recently.

This dissertation is motivated by this simple realization: scientific and technical metrics have become tethered with knowledge, yet these metrics are not given, but rather chosen —by imperfect humans often under imperfect information. Because we have become largely socialized to trust numbers in the realms of science and technology, these metrics are typically taken as given without much question or even attention. This dissertation aims to fill this gap, by studying the role of metrics in influencing which ideas get developed, how research and development resources get allocated, and ultimately, who benefits from innovative effort. This dissertation consists of three essays studying the role of metrics at different stages of the innovation process: the conception of using a metric for innovation purposes, the introduction of such a metric into an industry, and the consequences of when such a metric becomes entrenched.

In the first essay, I explore what defines and delineates the role of metrics in ideas production from other settings with less uncertainty. Prior work in the social sciences is largely dichotomous is its assessment of metrics, either emphasizing the benefits of metrics in facilitating comparisons and validating new knowledge, or warning that metrics eliminate the richness of context and homogenize effort. Because these studies draw upon a wide range of metrics and settings, it is challenging to interpret these contrasting perspectives. This paper synthesizes prior literature, identifying two key dimensions that metrics can vary along in how they are utilized. I then use these two dimensions as building blocks for a proposed organizing framework for understanding the different types of mechanisms that metrics can have. This is then used to structure a discussion of the ways in which metrics can be applied to different junctures in the innovation process.

In the second essay, I hone in on one of these junctures, when a new metric is introduced to an industry. Specifically, I study the impact of the Side Impact Dummy (SID) on US automotive safety innovation. SID was developed in response to growing policy and consumer demand for improvements in automotive safety, but for cost minimization and convenience purposes, was based on parts used for military dummies. This resulted in SID being modeled on the 1960s US average male, despite being introduced over three decades later. My data indicate that the introduction of SID as a metric led to meaningful changes in safety outcomes but also to disproportionate improvements for drivers similar to the metric. Furthermore, I find that this is moderated by different responses from firms depending on their prior strategic investments in safety so while metrics can accelerate innovation, they can also overly focus the attention of some firms in the narrow region of what is being measured.

Finally, in the third essay, I examine metrics at the other end of the innovation process when they have become widely accepted as the dominant metric. In particular, I seek to understand whether such metrics can be displaced when they have stopped serving a useful purpose and are generating negative externalities. Specifically, I study the impact of a EU policy intervention designed to limit the incentives of building upon watts through a ‘cap’ in the domestic vacuum cleaner industry. Although policymakers intended for the cap to shift innovative effort towards improvements in energy efficiency, I find that there was no discernible change in this direction. Instead, results show that rather than achieving the latter, there was a severe crowding out effect, with firms reducing their R&D in the focal area affected by the policy and channeling their efforts towards adjacent, unregulated product areas. Ultimately, this paper shows that dominant metrics can become entrenched in the R&D approaches of firms, and that a straightforward cap will not necessarily induce meaningful change, and opens questions for future experiments and research on more effective solutions.

Taken together, these essays form the beginning of a broader agenda to understand how scientific and technical metrics can impact innovation outcomes. Metrics have evolved to underpin modern knowledge, not only in the domains of natural sciences, but other areas of science, technology, social sciences and more. While they are certainly ubiquitous and treated as important, metrics are not given by Nature. Ddespite our high levels of trust and elevation of them as arbiters of knowledge, metrics are susceptible to our foibles and limitations. Understanding how metrics are developed and diffuse, how they operate in different phases of ideas production, and the consequences of their reification are therefore important questions that this dissertation only begins to pull the thread on and hopes future work will continue to uncover.

Chapter 2

What Do Metrics Do? A Review and Proposed Typology

Abstract

Metrics focus attention, and attention is typically scarce within modern life and organizations. Prior work is largely dichotomous in its treatment of metrics. On the one hand, some emphasize the benefits of metrics in augmenting human cognition, facilitating comparisons, and validating knowledge. On the other, others warn that metrics eliminate the richness of context, applies homogenization pressures and can distort their measurand in unintended ways. Because these studies draw upon a wide range of metrics operating in different settings, it is challenging to interpret these contrasting perspectives. Furthermore, despite their prevalence in the laboratory and in R&D departments, the role of metrics in scientific and technological innovation is largely overlooked in prior work. This paper aims to grapple with these two gaps in the literature by synthesizing extant work in management, sociology and economics to identify two dimensions that metrics can vary along in their use. Intersecting these yields a simple typology of four mechanisms that metrics can take on, and highlights when metrics might deviate from their intended purpose. This framework also allows for a structured exploration of the role of metrics in innovation.

2.1 Introduction

Consider an individual driving her children to school before heading to work herself. This mundane sequence of tasks actually involves several choices that are influenced by metrics. The vehicle may have been purchased for its high safety performance in crash tests (Liu et al., 2020), the school district may have been selected for its above average test scores (Gibbons et al., 2013; Hasan & Kumar, 2019), and the driving route may have been determined by a software application optimizing for minimal travel time (Graaf, 2018; Vasserman et al., 2015). This vignette illustrates that metrics have become enmeshed in modern life from large decisions to everyday minutia. This has only increased in relevance in recent years as the exponential growth of software has made it possible to measure more aspects of our lives, on an increasingly granular and frequent basis (Kellogg et al., 2020; Burrell & Fourcade, 2021). While this phenomenon has drawn attention from both social scientists and practitioners, prior work is largely dichotomous, either emphasizing the benefits of metrics in facilitating comparisons and validating new knowledge (Taylor, 1911; Drucker, 1965; Hubbard, 2000), or warning that metrics eliminate the richness of context and lead to homogenization (Espeland, 1993; Strathern, 1997, 2000; Muller & Muller, 2018). Because these studies draw upon a wide range of metrics and settings —ranging from corporate firms to the military to hospitals to academia—it is challenging to interpret these contrasting perspectives.

Furthermore, while prior studies have explored different aspects of metrics in modern life and work, it has left one area relatively unexplored: the role that metrics can have in shaping innovation. Metrics can provide an objective way to compare outcomes, and to validate those comparisons through replication. In other words, metrics are tightly entwined with the scientific method that scientific and technological norms. Coupled with the cumulative nature of innovation, this implies that metrics can influence the direction that science or technology progresses, often for sustained periods of time. For instance, continuing from the opening example, metrics used for crash tests can influence the research and development process for vehicular safety innovation, affecting the types of vehicles that ultimately become available on the market (Wu, 2022). Similarly, the types of education technology software that get developed will be determined by the metrics used to evaluate their efficacy, and routing software engineers will prioritize travel time over innovating along other, noisier dimensions such as neighborhood safety and environmental emissions.

The objective of this paper is to grapple with these two gaps in the extant literature —the disparate and contrasting perspectives on the impact of metrics, as well as the interplay between metrics and innovation. To start, in Section 2.2, this paper synthesizes related but siloed work across management, sociology and economics on metrics. This identifies two underlying dimensions that metrics vary upon. These two dimensions are intersected in Section 2.3 to give rise to four distinct mechanisms that metrics can play within organizations. In Section 2.5 this paper then uses this proposed typology as an anchor in exploring the influence of metrics in different phases of the innovation process. This highlights the role of scientific and technical metrics in the junctures of exploration and exploitation, leading to several open questions of future work. Section 2.6 concludes.

2.2 Towards a Typology

To ground ideas, metrics are provisionally defined in this paper as a method of mapping an object into a lower-dimensional representation. This definition encompasses both a measurement approach (a system of transformation the object interacts with) and the result that such an approach yields (typically a number but also a symbol or category) (Stevens, 1946). Therefore, this paper uses metrics to encompass cognate terms across different literature such as quantification, commensuration, and measurement. Metrics therefore involve condensing the amount of information about an object into a subset of dimensions to focus upon. This process lowers the cost of comparison through simplification of the objects and prioritization of what to focus on. This also makes it easier to validate these comparisons through replication, and to document and communicate these comparisons across distances in geography, knowledge domain and even time.

Prior scholars studying the effect of metrics on individuals and organizations have, by and large, focused on either the benefits that metrics can provide through simplification ("metric optimists"), or the negative consequences of that process ("metric pessimists"). Metric optimists tend to follow the canonical Taylor (1911); Drucker (1965), and place emphasis on the value of metrics in increasing transparency (Holmstrom & Milgrom, 1991; Gibbons et al., 2013), establishing trust (Crosby, 1997), and enabling optimization. By reducing the complexity of objects or agents, metrics can also augment limited human cognition and increase rationality (Zelizer, 1989; Kellogg et al., 2020), making it possible to increase managerial scope (Hubbard, 2000; Ichniowski & Shaw, 2003). Breaking down complex concepts into simple metrics makes it easier to understand and analyze the phenomenon (Martin, 1868; Weyl, 1947; Kleinert, 2009). This also allows for a shift from experience-based knowledge —in which individuals with intellectual authority have amassed reputation and access to large quantities of experience —to experiments-based (Lin, 1995), which lowers the barriers to entry for contributing ideas. Metric pessimists instead warn that metrics are often granted significant and unfounded levels of trust and objectivity by individuals and organizations (Porter, 1995). This can lead to negative externalities as metrics are found to alter incentives, engender strategic gaming, and lead to deviations from the intended objective (Kerr, 1975; Sauder & Espeland, 2009; Muller & Muller, 2018; Baker & Hubbard, 2003). Furthermore, metrics eliminate the richness of context. This can homogenize ex ante diverse objects, individuals or organizations into one (Espeland, 1993; Espeland & Sauder, 2007; Berman & Hirschman, 2018). Metrics are also cautioned as they allow for control and discipline at a larger and fine-grained scale, which can curb experimentation and autonomy (Foucault, 1977; Espeland et al., 2016; Mazmanian & Beckman, 2018).

Across these studies, a common theme is that the impact of a given metric will depend on its *application*. Yet with only a few recent exceptions (Christin, 2018; Ranganathan & Benson, 2020), prior work has largely bundled metrics and their applications together, rather than considering them separately. In fact, a single metric can serve heterogeneous roles in different application contexts. To unpack this further, the following subsections will focus on the different types of problems and intended objectives that metrics can be used for.

2.2.1 Known versus Unknown Problems

One dimension that metrics can differ upon in their application is the degree of structure of the problem or task at hand. Some metrics will be applied towards problems that have little to no rules or references, while others will be used for tasks that are clear and defined. While problems will exist along a continuum, a useful way to approach this is to draw upon the seminal classification of problems as being well-structured versus ill-structured (Simon, 1973).

A well-structured problem will have (1) a clear initial state, (2) a clear desired end state, and (3) defined and legal moves between them. An individual, organization or firm (herein, agent) facing the problem is aware of and can access full information about each of these three facets. These well-structured problems, referred to as tasks in some fields, have an optimal path for agents to follow, making the focus of the analysis on providing sufficient incentives for agents to exert effort towards that path, rather than which steps to take. For example, in canonical agency models such as Holmstrom (1979) and Holmstrom & Milgrom (1991), the agent has full information about the task(s) at hand, and the choice is the degree of effort allocation (between a single task and shirking, or between multiple tasks).

In contrast, ambiguous or ill-structured problems will lack one or more facets of the problem space. Sometimes referred to as the "residuals" of what well-structured problems are not (Simon, 1973; Reed, 2016), these types of problems have multiple solution paths as well as uncertainty surrounding the rules, constraints and concepts that matter for the solution (Johanssen, 1997). This can be due to limited information (Holmstrom 1989), conflicting or irrelevant information, or fundamental uncertainty (Aghion and Tirole, 1994; Knight, 1921). For example, in models of agents under uncertainty, agents have to choose between multiple divergent, potentially irreversible paths without full information on which one is optimal (Manso, 2011; Gans et al, 2019).

Problems typically do not fall cleanly into one category or another, but rather somewhere on a continuum of problem structure. With the caveat that this will depend on the agent, examples of problems that fall on the well-structured side of the continuum are solving a quadratic equation, assembling a piece of furniture, playing a game of chess, or solving a cryptarithmetic puzzle (Goel, 1992). In such settings, metrics are typically used to allow principals and agents themselves to better observe the output of a given problem. Examples on the other end of the continuum include designing a piece of furniture, solving a crisis, or mitigating the effects of a pandemic (Reed, 2016). Under these types of conditions, the agent not only needs to grapple with how much effort to apply to a given problem, but also how to define the core of the problem in the first place. Metrics in this instance, are often used not only to observe output, but also as a way to add structure and define the problem.

Understanding the implications of a metric on agents' choices and behavior therefore, requires consideration of the degree of structure that a problem has. Problems can also, over time, move from being ill-structured to well-structured (e.g. additional information about a problem space is revealed as search effort increases), and vice versa (e.g. a shock disrupts the previously stable rules governing the solution). This means that a metric applied to one problem can still take on different mechanisms as the underlying problem evolves.

2.2.2 Signal versus Coordination

Another dimension that metrics can vary along in their application is the intended objective for developing or using the metric. In other words, this is the planned function that a given metric will play in the market (Bothner et al, 2022). This will depend on the level of information available, and the ways in which agents affect one another.

On the one hand, metrics can serve as information mechanisms by reducing noise and signaling progress on a problem. In settings where there is asymmetric information and noise, it may be challenging to discriminate among agents, objects or ideas. Metrics, because they are universal, non-rivalrous in their application and relatively more objective than qualitative information, can be perceived as credible ways to understand the relationship between the resulting outcome and the inputs (e.g. ability, effort, fit) (Holmstrom & Milgrom, 1991). For example, firms often have multiple candidate objects (individiduals or ideas) that appear equally promising. This can be mitigated by using a metric that can distinguish between candidates that perform better versus poorer than others (e.g. through a pre-employment test, years of education) (Spence, 1978). Relatedly, firms could be assessing different technologies that each have their own advocates and introducing a metric (e.g. performance on a benchmark test) could inform decision-making (Vinokurova & Kapoor, 2020). This reduction in noise between the initial inputs and the resulting outcome means that an agent can more clearly distinguish themselves from competitors, or a principal might be better able to identify progress, shirking and bottlenecks more easily.

On the other hand, metrics can act as coordination mechanisms by lowering the costs of aligning different objects or agents with each other. In settings where there are frictions to the adoption of a new idea (e.g. due to interdependencies, high fixed costs), or agents have uncertainty towards one anothers' incentives, it can be challenging to coordinate even when it is optimal to do so. Metrics, because they are easy to communicate and non-rivalrous and transparent in their usage, offer an effective means for communicating and enforcing a common objective or norm by identifying deviations efficiently. Moreover, relative to a setting without metrics, agents can coordinate with a greater number of parties, and to do so across domain, time and geographic distance. In contrast to signaling, in which a metric creates value by allowing agents to distance themselves from one another, in such instances, a metric allows agents to minimize that distance and ensure compatibility with an agreed upon threshold. For instance, firms may benefit from product interoperability, but due to strategic uncertainty and frictions in assessing one anothers' products, there is reluctance to invest the costs to modify their existing products to meet a standard (David and Greenstein, 1990; Hemenway, 1975). Introducing a metric can reduce the communication costs between firms by creating a common "language" as they come to agreement, and then assess one anothers' compliance.

Taken together, metrics do not exist in a vacuum, but rather can be applied towards a wide range of problems and objectives. This next section will intersect these two dimensions of application to propose four different types of mechanisms that metrics can exhibit.

2.3 Four Quadrants

A given metric can enact distinct mechanisms depending on how and where it is applied. Even if developed to solve a specific problem, a metric can be employed towards alternative, unplanned use cases. To make this concrete, consider dry matter content, a metric that measures the soluble fibers and starches that remain in food after water is fully extracted. The initial key use case for dry matter content (here in DMC) was to assess the water content of livestock feed for herd health, but has recently been applied towards several problems in fruit production. First, DMC has been used to rank different pruning methods against one another in terms of the quality of fruit that is produced (Goke et al., 2020). It can also be used to ensure that fruit meet a certain quality threshold when grown at farms under variable environmental conditions (Chagné et al., 2014). DMC also offers scientists a way to jointly explore novel prediction models of fruit yield (Maeda & Ahn, 2021). Finally, DMC has been employed as a *target* for engineers developing new varieties of fruit with wide consumer appeal in taste (Kaczmarska et al., 2016). This example of DMC shows that single metric can exhibit contrasting mechanisms with some use cases that require distinguishing between different objects, and others that prioritize homogeneity; as well as some use cases that deal with relatively known methods, and others that involve exploring nascent technologies.

Figure 2-1: A simple typology of metric mechanisms

	Signal	Coordination
Known	**Rank**	**Normalize**
Unknown	**Target**	**Joint Search**

In an effort to organize these different mechanisms, the two dimensions that metrics vary upon in their application —*known* versus *unknown* problems, and *signaling* versus *coordination* —can be intersected to yield a simple typology. Figure 2-1 classifies four types ways in which metrics can be utilized: (1) to produce a rank ordering, (2) to normalize, (3) to conduct joint search, and (4) to set a target.

2.3.1 Metrics for Ranking

In the upper left quadrant, metrics can be used to produce a ranking of different agents, objects or ideas against one another. In such settings, the problem is *known* and well structured, and the objective for using the metric is to reduce noise and lower the costs of comparison. On the one hand, this can generate a better signal of which agent, object or idea is better than others along the dimension of the metric, making it easier to choose how to allocate resources (Holmstrom & Milgrom, 1991). This can also improve the visibility that a principal or even agents themselves have on their relative performance, which can increase incentives for improvements (Baker & Hubbard, 2003). At the same time, rankings enhance surveillance and control capabilities, which can reduce variation and autonomy (Berman & Hirschman, 2018). This consequence is further exacerbated by the fact that a ranking necessitates prioritizing some dimensions over others, abstracting away potentially valuable heterogeneity. In addition to the example of using dry matter content to rank different pruning methods against one another, other examples of metrics that result in rankings include school ratings (Hasan & Kumar, 2019), using RFID tags to measure employees' garment production productivity (Ranganathan & Benson, 2020), Yelp scores to gauge the quality of a restaurant (Luca, 2016) and news media click-through rates to assess the success of an article (Christin, 2018).

2.3.2 Metrics for Normalizing

In the upper right quadrant, metrics can be used to enforce a norm across diverse agents. As with rank metrics, these types of metrics are applied to known problems. In contrast however, the focus is not on producing a rank order, but rather on coordinating across multiple agents or objects. Metrics lower the cost of communication, allowing a common understanding to be achieved more efficiently. Furthermore, by making comparisons easier to replicate, they also lower the costs of monitoring and auditing for compliance with the agreed upon measures. This allows for normalization across a greater number of agents or objects, over organizational, domain or geographic boundaries. Normalization allows for interchangeability, which can generate positive network externalities such as scale economies in production (Hemenway, 1975) and supplier networks (Basker & Simcoe, 2021). At the same time, such metrics can generate increasing returns to adoption that ‘lock in’ agents and objects onto a suboptimal metric (David, 1985; David & Greenstein, 1990). The dry matter example of an application for normalization is using the metric to ensure a consistent quality across different farms. Other examples of these types of metrics include the number of violations prohibited in food safety regulations (Ibanez & Toffel, 2020), weight thresholds of newborns in clinical guidelines (Almond et al., 2010), and the major diameter and pitch in screw thread standards (Hemenway, 1975; Yates & Murphy, 2019).

2.3.3 Metrics for Joint Search

In the bottom right quadrant, there are metrics that unite agents or objects together for a joint search for advances on a problem. In contrast to metrics used to normalize, problems in such settings are generally ill structured with many facets of the solution space that are *unknown*. Metrics allow for a *coordinated* approach to the problem by providing an anchor to explore the problem space, and providing a definition of what meaningful progress means to the broader community or field. Metrics in such applications create a common language for discussing and evaluating different ideas, which not only lowers redundant or incompatible efforts, but can also facilitate an improved exchange of ideas (Arora & Gambardella, 1994) and learning spillovers. In addition, by providing a common cognitive frame for an ill structure problem, this makes it easier to establish trust under uncertainty (Gibbons et al., 2021), shifting the need for different individuals or organizations to trust one another towards trust in the metric. Metrics can thus enable collaborations that previously would not have occurred such as those without prior history, or operating in nascent markets. This is similar to the concept of gestalt, in which a simplified and holistic approach can facilitate sense-making in nascent problems, and allow the whole is able to achieve more than the sum of the individual agents' efforts (Rindova et al., 2010). Dry matter being used as a novel metric for modelling fruit yield is one example of a metric being used for joint exploration. Other examples of these metrics include the use of millisecond latency to create consensus in 5G Ultra-Reliable Low-Latency Communication (Gupta et al., 2019), the astronomical unit (au) in astronomy as a metric that harmonized the field in how to quantify astronomical distances (Luque & Ballesteros, 2019), and cases per 100 thousand people as a unified policy approach to deal with COVID-19 containment for public health.

2.3.4 Metrics as a Target

In the bottom left quadrant, metrics can provide a target for agents. Similar to joint search metrics, these are settings where there is a problem that has many elements that are uncertain and *unknown*. In contrast however, rather than focusing on establishing a coordinated approach to solving the problem, the objective of such metrics is to offer a meaningful target for agents, objects and ideas to work towards. In these applications, advancing along a metric allows agents to signal their success in pushing out the scientific or technological frontier, which can attract greater effort and talent to solving the problem. Metrics shift progress on an ill-structured problem from being driven by "soft" information to "hard," lowering the barriers to entry for novices and outsiders by substituting reputational gatekeeping with performance on the metric. At the same time, although these metrics are chosen as targets under limited information, they have a tendency to persist because they are trusted, objective norms (Porter, 1995). Returning to the dry matter example, an application where it serves as a target is in the genetic engineering of new fruit varietals. Other examples of target metrics include thoracic trauma as a metric for vehicle manufacturers to focus on lowering (Wu, 2022), floating operating points per second in supercomputers (Dongarra et al., 1979; Dongarra, 2006), progression free survival in oncology clinical trials of therapeutics (Booth & Eisenhauer, 2012), and emissions ratings of firms (Chatterji & Toffel, 2010; Sharkey & Bromley, 2015).

2.4 Framing Metric Mishaps

Putting metrics into these four quadrants of potential mechanisms helps clarify the dynamics of metrics, and the consequences that can arise in their usage. These four applications of metrics —to rank, normalize, jointly search and target —are not necessarily static. A metric applied to a given problem may over time, shift from one quadrant to another depending on the dynamics of the problem. For instance, as additional effort is applied to a problem over time, much of the uncertainty may be reduced such that the problem becomes more structured. In such cases, what might have start out as a "joint search" or "target" metric might become applied as a "normalize" or "rank" metric instead.

Issues or challenges however can arise, as these quadrants can be rather porous in their interpretation, and a metric developed or chosen for one mechanism may unintentionally end up being utilized as another. While the four mechanisms can mathematically lead to several potential 'misuses', the two that are most common in prior work are (1) metrics that were intended for ranking individuals, organizations or objects that become targets, (2) those intended for ranking that become used to normalize.

2.4.1 Rank to Target

Metrics intended for ranking assume that the structure of the solution is known and stable (i.e. there is a consistent set of steps from the initial state to the desired end state), and that the individuals, objects or organizations being measured are immune to the knowledge of the metric. For instance, if one were to evaluate different schools by introducing a metric for ranking them, a key assumption is that the educational practices, hiring decisions and admissions policies of the school will remains stable regardless of the use of the metric.

Sociologists have highlighted that this is inconsistent with many settings in practice. In fact, metrics will often engender a behavior labelled 'reactivity' in which the individuals or organizations being measured will change their behavior to accord with the metric (Espeland & Stevens, 2008; Espeland et al., 2016). Critiques of a similar vein from economics and accounting include the Lucas Critique (Brunner et al., 1983) and Goodhart's Law (Goodhart, 1984; Strathern, 1997). This reactive desire to advance along the metric could be driven by financial incentives, status rewards, identity or discipline avoidance (Ranganathan & Benson, 2020). Expressed using the framework, these are metrics that were intended to be used for ranking that end up becoming targets for those that are being measured. For instance, the intention of introducing a school ranking may have been to increase accountability and encourage the faculty and leadership at schools that were falling behind to apply more effort towards their educational practices. Instead, what might occur is that even high performing schools might be destabilized by the metric, and induced to rethink their well-structured pedagogical approach in efforts to maximize their position on the new metric. Other examples include viewership metrics in YouTube for helping creators understand consumer demand that end up becoming treated as a sign of social status (Christin & Lewis, 2021); or click rates in news media that cause some journalists to change the articles they pursue towards “click bait” (Christin, 2018). Therefore, the metric that was intended to gauge the performance of individuals or organizations, will end up affecting what performance even means in the first place (Kerr, 1975; Mau, 2019).

2.4.2 Rank to Normalize

In addition to the above, metrics that are applied to produce a ranking can also have another unintended reactive consequence. The process of transforming multidimensional individuals, organizations, and objects into a metric (i.e. commensuration) involves prioritizing certain dimensions over others. This creates information loss which can erase contextual features, eliminate authenticity (Zelizer, 1989) and reify one dimension as the ‘gold standard’ (Espeland, 1993; Fourcade, 2011). This can create pressures to conform with what is being measured, losing *ex ante* heterogeneity in the process. For example, as reviewed in Espeland et al. (2016), school leaders can start to mold their organizations towards the metric such that heterogeneous goals get lost. This means that whereas some schools may have admitted non-traditional students before, the homogenizing pressures introduced by the metric eliminate opportunities for such students. Whereas the introduction of the metric was intended to distinguish between different individuals, organizations or objects, the interpretation of the metric as the ideal, can have the opposite effect of normalizing them to be the same.

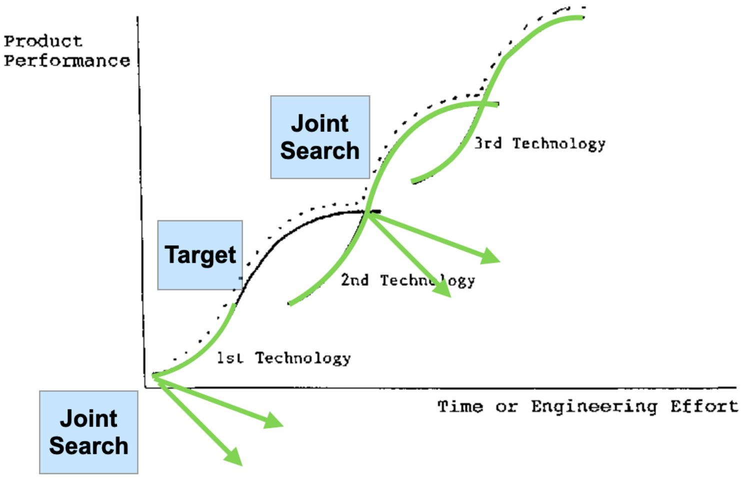
2.5 Metrics in Innovation

While scientific and technological innovation will necessitate the use of metrics for known, routine tasks —for instance to calibrate a microscope using micrometers, or ranking tools based on their precision tolerance —this next section will abstract away from these functions, and focus on the role of metrics in unknown and uncertain problem spaces. The scientific and technological innovation process involves periods of stable development interspersed with periods of revolution as new paradigms displace the old (Kuhn, 1962; Dosi, 1982; Henderson & Clark, 1990; Christensen, 1997). These distinct phases can be characterized using the canonical exploration versus exploitation framing (March, 1991). During phases of exploration, there is a wide search within the uncertain problem space, with relatively high risk-taking and experimentation as scientists and engineers search for the optimal solution. During phases of exploitation, this becomes more focused on refinement, improvement and execution along a particular solution path (March, 1991; Levinthal & March, 1981). Metrics play important roles at the junctures between these two phases.

2.5.1 Exploration to Exploitation

During phases of exploration, there are often multiple, potential paths forward that can get surfaced. Because these paths are equally attractive and involve uncertainty, it can lead to inertia. Resolving uncertainty to understand which path to pursue would necessitate making commitments towards one over the others, which could be an irreversible action (Gans et al., 2019). This suggests that there can be a point where the nature of exploration reaches a stalling point for making progress on the problem at hand. Furthermore, multiple individuals may pursue different paths with no means to adjudicate which one is more promising to invest in.

Figure 2-2: Metrics throughout the innovation process



Metrics can be applied to mitigate this problem by using them as a *joint search* mechanism. Metrics provide a heuristic that decomposes a complex and uncertain problem into several smaller sub-problems. This requires the organization or field to come together to clarify what is considered a desired direction of progress, and can bound the initial search space. Providing a joint search metric also allows for different paths (and scientists' work) to be assessed using this mutually agreed upon metric to collect data. Metrics are a core input into the scientific method, allowing for hypotheses to be expressed in an experimental way such that evidence can be gathered. Put more eloquently, metrics provide "a quantitative representation of your subject, however simplified, even in its errors and omissions, precise. You can think about it rigorously. You can manipulate it and experiment with it" (Latour, 1986; Crosby, 1997).

2.5.2 Exploitation to Exploration

As uncertainty about a problem space is resolved through the use of joint search metrics, a particular path tends to emerge as the scientific or technological frontier. To exploit this path, a metric can be used as a *target* to encourage advancements that push out the frontier. Because the path forward is relatively defined, metrics in such instances, can propel improvements by providing a means for benchmarking progress. Even without pecuniary incentives attached, this reward of being at the frontier or top of ‘leader board’ can attract increased effort and participation.

At the same time, metrics can not only encourage the exploitation of a given path, but can also lead to discoveries that lead to the transition to the exploration stage. Much like a compulsive sequence (Rosenberg, 1969), advancing towards a target requires gathering data and conducting experiments. This can lead to incongruencies in which stylized evidence does not accord with or is poorly captured by existing metrics. This can move innovation into a phase of exploration, to understand how to reconcile prior theories or develop new ones, and potentially to develop new metrics as well. Kuhn (1961) characterized this as metrics fomenting “scientific crises” that can lead to new paradigms.”

2.5.3 Known Unknowns?

Although metrics in innovation are generally operating within uncertain and ambiguous problem domains, they often become treated as if they are known facets of Nature or some other ground truth. This is because, as Thomas Kuhn observed, most people first encounter innovation metrics in a science textbook or lecture hall Kuhn (1961). These pedagogical settings generally frame metrics in scientific and technological innovation as objective, law-like truths. As science historian Porter

(1995) explains, within modern science norms "measurement means nothing if not precision and objectivity." This implies that metrics have the potential to define the knowledge space itself. For instance, in a traditional technology S-curve plot, the Y-axis used for assessing progress on the problem endogenously reflects the metric that was initially selected. This suggests the choice of metric is crucial for individuals and organizations, as it can come to be the dominant lens for thinking about a problem and may not be challenged as readily in the future.[1] Specifically, a metric can exhibit path dependence due to scale economies where the marginal benefit for adopting a metric increases along with the total number of inventors that utilize it, as well as the number of complementary measurement tools or skills that get developed (David (1985). Metrics can end up exhibiting similar characteristics to that of a dominant design as they become widely accepted as characteristics of a given knowledge space (Utterback & Abernathy (1975); Anderson & Tushman (1990); Utterback (1994)).

To draw upon the four quadrants, the concern in these cases is that a metric intended for 'joint search' of an unfamiliar problem space may end up prematurely being treated as if it were a metric intended to 'normalize.' Rather than recognizing that this was an *ex ante* specified metric for guiding innovation, it can become treated as an approach that must be conformed with, stifling the experimental activity it was meant to unleash. Similarly for metrics that were intended to serve as 'targets' being treated as 'rank' metrics before uncertainty has been resolved. Rather than

[1] This also has the implication that metrics can have strategic importance, in terms of whether and how to disclose a metric to the market. To clarify with anecdotes from practice, firms have been observed to take different approaches to innovation metrics. On the one hand, some firms such as Intel share a metric (transistor count) widely with their competitors (Moore's Law). Intel may have found it advantageous to shape innovation in the domain to be in line with its own strategic advantage. In contrast, other firms such as Bose have intentionally rejected using industry metrics and have kept their own metrics proprietary. This may be because revealing the metric could disclose to others how Bose is approaching the problem of improving audio quality, and could lead to competitive imitation. This suggests that metrics can be innovations in and of themselves.

encouraging experimentation and risk taking to progress along the target, scientists or engineers could be dissuaded from such high variance activities out of fear that failure would disrupt their rank order (Manso, 2011; Azoulay et al., 2011). Taken together, metrics can persist within a domain, even if alternative metrics are proposed or the metric becomes obsolete. This is important to highlight because given that they are developed or chosen to grapple with unknown problem spaces, it is limited and endogenous to what a particular scientist, engineer, manager or even policymaker might be familiar with at a given point in time. Rather than being revised as more information about a problem becomes available, these metrics can end up being counterproductive as they get treated as facts of knowledge rather than the best guesses for confronting the unknown.

2.6 Conclusion

"What gets measured, gets managed" is a commonly expressed piece of management wisdom: we can only improve the performance of what we measure, and choosing what to measure is a key strategic decision that can deeply shape the performance of individuals, organizations and institutions. This paper suggests that this popular adage captures only one mechanism of several that metrics can enact. By synthesizing prior work studying a range of metrics in different settings, this paper shows that metrics can take on different mechanisms depending on their application context. A given metric can be used in addressing problems that have varying levels of structure, ranging from highly structured and well-known tasks to ambiguous, nascent problems that are open-ended. They can also be utilized to achieve different objectives, from using them as a means to differentiate between the individuals, objects or organizations being measured to the other end of the spectrum, where the goal is to minimize any discrepancies. Putting these two dimensions together shows that beyond just being a managerial tool, there are four nuanced roles that metrics can play. Whereas they can certainly be used as a mechanism for *ranking* that increases visibility and identifies inefficiencies, they can also be used as a way to *normalize* to a certain standard or threshold, allowing for compatibility, economies of scale, and quality control. A relatively less explored area is when metrics are applied in contexts where uncertainty is high and there is the possibility to innovate along multiple diverging paths. In such settings, metrics are not only managerial tools, but can also define the overall idea space for *joint search* to occur within. Metrics can also provide a *target* for innovators to work towards, influencing not only the direction but also the rate of innovation. Framing metrics as having four potential mechanisms clarifies where the consequences outlined in prior work can arise (reviewed in Espeland et al. (2016), Muller & Muller (2018), and Mau (2019)). It also clarifies the role that metrics play during different phases of the innovation process, offering a means for overcoming inertia and redundancy through collaborative exploration, and a benchmark for propelling the frontier during phases of exploitation.

The simple typology proposed in this paper leaves many open questions for future work. While this paper explores two ways that metric slippage can occur that are common in the examples cited in extant work —when a metric intended to form a ranking becomes used as a target, or when it becomes interpreted as a standard to normalize to —there are several other cases of metric misuse that could occur. In addition, future work could focus on the factors can affect how a metric is interpreted relative to its intended use. Recently, Christin (2018) identified national cultural contexts, Ranganathan & Benson (2020) found task complexity, and Wu (2022) found prior domain knowledge to be moderators for how individuals and organizations react to metrics. Another fruitful direction is to further explore the role of metrics in innovation and knowledge, especially how they come to be created and gain traction. Metrics in such settings, can be an innovation in and of itself, and also one that can unlock further measurement. Practically, metrics can be traced through papers, patents, popular articles, syllabi, and in some fields, benchmark competitions serving as a paper trail for the study of innovation dynamics and the evolution of technological opportunity in a field (Cohen, 2010).

Ultimately, measurement is attention, and attention is typically scarce within modern life and organizations. On the one hand, metrics can augment human cognition, breaking down complex problems into penetrable and salient units. This allows for improved comparison, prediction, management, and collaboration. At the same time, metrics are not applied in a vacuum, and what may start as an objective definition of merit or desired norm can deviate away from intention as individuals or organizations react to a metric (intentionally or not). This can result in metrics distorting or exerting control over the very objects, individuals or organizations they were were intended to help. Understanding when and why the use of metrics can lead to one outcome over the other is an important area of significant scholarly, management and policy importance.

Chapter 3

Innovation for Dummies? Exploring the Role of Metrics in Automotive Safety

Abstract

Metrics permeate our daily lives, firm R&D, and scientific and technological progress —yet because we often take them as given, we have little understanding of their impact on innovation. This paper explores how metrics shape the rate and direction of innovation. By taking advantage of a unique change in the US automobile industry, I estimate how the introduction of a novel metric for automotive safety, the Side Impact Dummy (SID), influenced safety outcomes and firm performance. Using rich data on US vehicular accidents and car specifications, I find that the new metric significantly reduced fatalities, but did so disproportionately for occupants similar to SID in body size. These results were driven by two types of firms: while firms with pre-existing strategic investments and knowledge in safety made improvements that benefited everyone, those without a history in safety ended up narrowly focusing on the metric. Taken together, these findings show that metrics can profoundly affect the innovation trajectories of firms, and that firms without ex-ante domain expertise may suffer from "metric myopia." Moreover, the choices firms make in response to a metric can, consciously or not, determine who benefits from innovation.

3.1 Introduction

A long-distance colleague asks “how warm is it over there?” The response for most people would be to state the temperature in Celsius or Fahrenheit, rather than go into a lengthy description of individual sensations. After all, this is an effective and succinct way to convey information. Scientific and technological metrics are how we, as modern society, define what it means to *know* something (Kelvin, 1889).[1] As a result, these metrics permeate our lives —yet we often take them for granted. This extends to management research, where there has been surprisingly little work on how scientific and technological metrics impact firms and individuals. Across industries, metrics are generally viewed as law-like, objective representations of the truth, but this perception may need further scrutiny (Kuhn, 1961).[2] After all, metrics often stem from haphazard choices and idiosyncratic circumstances.[3] Moreover, once adopted, they can persist and influence the rate and direction of innovation for centuries. For instance, while developing his eponymous steam engine, James Watt introduced a metric to compare his invention against the main alternative at the time, factory horses (Scherer, 1965).[4] The resulting horsepower metric ended up

[1] As Lord Kelvin stated: “when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind” (Kelvin, 1889).

[2] As Thomas Kuhn stated: “our most prevalent notions about the function of [such] measurement and about the source of its special efficacy are derived largely from myth” (Kuhn, 1961).

[3] For example, skin color balance is a metric that was created at Kodak, and based on a random female Caucasian employee named Shirley out of convenience. This metric went on to shape film innovation such that many decades later, film photography did a poor job of capturing people of color (Roth, 2009). As another example, a Xerox engineer hastily created a test document as a metric for evaluating different printer prototypes during an “informal demonstration of laser printers for [executive] taskforce” which went on to “redefine the meaning of computer printing” (Vinokurova & Kapoor, 2020).

[4] James Watt was responsible for the technical breakthroughs and the actual invention of the steam engine, but his business partner, Matthew Boulton supplied the insight that a commercially attractive opportunity would be to apply it as a replacement for draft horses in manufacturing factories (Scherer, 1965).

defining progress and forming the basis of competition among firms not only for steam engines, but also for trains, turbines, automobiles, and more;[5] outliving the use of factory horses by centuries.

This paper is a first step towards understanding the role metrics play within the innovation process of firms. At a fundamental level, metrics lower the costs of comparing objects against one another. This however, is achieved by simplifying an object into a few key dimensions, and therefore, involves abstracting away some information and context. While prior work has highlighted how such loss of information may distort incentives for individuals or organizations, this paper surfaces how in the innovation process, this is also a valuable feature. In innovation, metrics help reduce uncertainty by allowing for a concrete definition of progress (whether it is optimal or not). The use of a metric can therefore accelerate the innovative output of firms by giving them a clear target to align their resources and R&D efforts towards. Metrics, by helping to quantify progress, are also a crucial input into the scientific method where they are used to set baselines, evaluate experimental outcomes against each other, and create a common understanding.

The entwinement of metrics and the scientific method within R&D can lead to the perception of metrics as fact. Firms and individuals may therefore overlook the subjective choices that are involved in the early development of a metric, and solely focus on it without considering the broader problem underlying it. Taken together, metrics present a trade-off to firms: on the one hand, metrics can summarize complex, uncertain problems into clear targets. On the other hand, their reification can lead

[5] Horsepower is not unique, as metrics exist across a number of industry contexts and shape innovation within them. Consider for instance, FLOPS driving innovation within supercomputers (Dongarra et al., 1979; Dongarra, 2006), watts within domestic vacuum cleaners (Wu, 2021), and progression-free survival as a metric within cancer clinical trials (Booth & Eisenhauer, 2012; Prasad et al., 2015)

to an overly narrow focus on the metrics themselves.

Understanding how the benefit of focusing innovation with a target weighs against the cost of potentially influencing it in a subjective direction is an important empirical question. From an inference perspective however, it is often impossible to credibly identify alternative ways a metric could have been defined. So while horsepower clearly propelled the development of increasingly powerful steam engines, it is difficult to know what other valuable dimensions of progress were ignored as a result. Furthermore, a firm's choice to adopt a metric is not necessarily random, which makes it challenging to separate out the impact of a metric from a firm's selection into adopting it. An ideal setting would therefore, offer both an exogenous introduction of a metric to firms, as well as involve a metric with a clearly defined set of alternatives.

The introduction of the Side Impact Dummy (SID) as a new metric for automotive safety not only provides the right variation to estimate the causal impact of the introduction of a new metric on firm performance and innovation, but also does so for outcomes that are economically and socially significant.[6] SID was developed by the U.S. Department of Transportation (DOT) and introduced in 1994 to automotive firms as a metric for occupant protection in side impact crashes. At the time, vehicle models were at different stages in their R&D life cycle: while some models had to wait for their next major redesign to incorporate SID, others could incorporate the new metric immediately into their R&D process. The setting therefore provides quasi-experimental variation in the timing of adoption of a new metric. Furthermore, although there was a known distribution of body sizes for the US population, SID could only be modeled after one body type. For cost and convenience purposes, the DOT re-purposed parts from older crash dummies to build SID, resulting in

[6] The automobile industry is estimated to account for over 3-5% of global economic output.

it inheriting the height and weight of the average US male in the 1960s (5'8" and 169 lbs). This setting therefore is unique in providing a bounded set of alternative body sizes SID could have been based upon, and variation in how representative the selected metric was for different vehicle occupants.

I rely on these two sources of variation in a differences-in-differences framework to estimate the impact of SID on automotive safety and firm innovation. Using rich data covering US vehicular accidents, I leverage the variation in the timing of adoption of SID in different vehicle models to examine the impact of the metric on safety outcomes. Because these vehicles were on the road and involved in similar types of crashes at the same time, this specification allows me to control for differences in safety between different vehicle models through vehicle model fixed effects, in the market over time through vehicle production year fixed-effects, and any firm-level shocks through firm and vehicle production year fixed-effects.

Results show that the new metric drastically reduced the likelihood of occupants dying in an accident by 48-50%. Using comprehensive car specifications data, I also find that firms made meaningful and costly changes to their vehicles in response to SID, increasing the width, length and curb weight to improve side protection. That said, not everyone benefited equally from these improvements. Using granular data on occupant characteristics, I find that occupants with a body type close to SID in height and weight were the ones with the greatest reduction in fatality risk.

These results suggest that introducing a metric can raise the overall level of innovation and performance within a problem space, but can also disproportionately channel effort towards improving along the definition of the metric itself.

When I explore outcomes at the firm level, I find that these aggregate results were largely driven by different types of firms. To understand firm heterogeneity, I collected data on patent applications and defects investigations to measure the level of pre-existing knowledge in safety firms had before the introduction of SID. I use patents to proxy for the safety R&D capabilities of the firm, and defects investigations to capture the ability of a firm to produce vehicles without safety issues once on the road.

Across both of these dimensions, two types of firm responses were visible in the data. While firms with prior domain knowledge in safety responded to the new metric by accelerating innovation benefiting all occupants, firms that were new to safety only improved outcomes for SID-like occupants. This suggests that firms with pre-existing knowledge may be less susceptible to "metric myopia" and realize that the metric involves information loss and captures only part of a problem. Possibly, when using the metric, they may be able to leverage their broader understanding of automotive safety to innovate in a way that is not constrained by the metric itself. Firms without that pre-existing knowledge instead, may use the metric as their only lens on the problem. This finding suggests that in the process of innovation, a metric is a *complement* to broader problem understanding rather than an actual substitute.

Lastly, in this setting, the metric was introduced by the DOT to all firms. This meant that both ex-ante safety-oriented firms as well as those that had not prioritized safety, suddenly had access to a common metric. This could have created a tension for firms with prior strategic commitments to safety as the new metric also allowed firms to be benchmarked more easily against one another. Ranking at the top could make their strategic advantage more salient in the market. At the same time, the metric provided all firms with a common lens for decomposing the problem, potentially lowering the barriers for imitation. In order to examine how this might affect the firms' responses to SID, I collected textual data from annual shareholder reports and media coverage of new vehicle launches. I applied machine learning tools to analyze the text for safety content, which allowed me to measure the pre-existing level of firms' strategic commitments to safety. Across both measures, I find that firms with a relatively lower safety orientation focused narrowly on SID, while those that had made safety a strategic priority focused on all occupants. This is consistent with safety-oriented firms being more cautious about fully embracing SID and continuing to maintain their own alternative metrics in automotive safety innovation.

Overall, these findings show that metrics exist not only within R&D departments, but also can significantly influence the broader innovation strategies of firms. Furthermore, it can generate a heterogeneous response across firms, depending on their level of prior knowledge and strategic commitments to the area of the metric. These findings offer contributions to three main areas of research.

First, it extends the work on performance measurement and its impact on firms and individuals (Kerr, 1975; Holmstrom & Milgrom, 1991; Espeland, 1993; Sauder & Espeland, 2009; Muller & Muller, 2018) by considering metrics in innovation, where tasks are not defined ex-ante, and there can be multiple paths forward. In such settings, individuals and firms do not necessarily react in the same way to a metric. This paper finds that prior knowledge and strategic commitments can be a mechanism that generates heterogeneous responses, and highlights the need for future study of how widespread this effect can be. Second, it contributes to the literature on standards (Bresnahan & Yao, 1985; Yao, 1988; Greenstein, 1993; Simcoe, 2012; Lerner & Tirole, 2014). Whereas prior work has focused on metrics as part of standards (e.g. as a reference or outcome standard), this paper surfaces that metrics can involve their own, unique mechanisms for innovation.[7] Last, I also expand work in innovation examining sources of technological opportunity (Rosenberg, 1969, 1982;

[7] This is especially relevant in settings when a standard or threshold is challenging to specify ex-ante, when compliance is costly to enforce, or when the objective is to lower the costs that firms face in innovating along a new dimension.

David, 1985; Farrell & Saloner, 1985; David & Greenstein, 1990; Arora & Gambardella, 1994). This paper contributes to this rich literature (see Cohen (2010) for a review) by proposing scientific and technological metrics as a mechanism that can induce and shape innovation. The setting of this paper also relates to research exploring the impact of regulatory and technological mechanisms on automotive safety outcomes (Grabowski et al., 2004; Levitt & Porter, 2001a,b; Dee, 1999; Graham, 1984; Peltzman, 1975), including recent work by Liu et al. (2020) on the effect of crashworthiness ratings design.

The layout of the remainder of the paper is as follows. Section 3.2 explores the role of metrics in the innovation process, and develops a conceptual framework of how such metrics can impact firms' innovation strategies. Section 3.3 describes the empirical setting of automotive safety. Section 3.4 outlines the data and empirical strategy. Section 3.5 summarizes the main findings from the causal estimates of the impact of introducing a metric on innovation in aggregate and at the firm level. Section 3.6 concludes.

3.2 Conceptual Framework

3.2.1 What is a Metric?

At a fundamental level, a metric summarizes an object into a lower dimensional representation. As a basic example, an apple, a pear and an orange are multi-dimensional objects that can be mapped onto a number of different metrics such as calories, mass, pH level, harvest time, and so on. I define a metric as encompassing both a measurement approach (i.e. a system that the object interacts with) and the resulting measure that such an approach yields (i.e. a number or category). To continue the example from above, calories as a metric refers to both the Atwater system and the numeric value of calories generated by the system.

Use of a metric, therefore, can significantly lower the costs of comparison, as well as validating and communicating the outcomes of such comparisons. At the same time however, by construction, the development of a metric will involve some loss of information and context.

Recent research, perhaps in response to widespread adoption of Taylorism and scientific management (Taylor, 1911; Drucker, 1965), has largely focused on the costs that loss of information can create when metrics are used. Across management, sociology and economics, scholars have raised concerns about metrics as homogenization mechanisms that can eliminate valuable context when transforming objects into comparable numbers. Because a metric rewards effort with a clear signal, it can elicit individuals or firms to engage in strategic behavior to advance along it (Espeland, 1993; Berman & Hirschman, 2018; Espeland & Sauder, 2007; Kerr, 1975; Goodhart, 1984; Strathern, 1997; Sauder & Espeland, 2009; Mazmanian & Beckman, 2018; Espeland et al., 2016). Individuals and firms will be inclined to over-invest in the metric, and therefore, managers and policymakers need to use low-powered incentives to combat this tendency (Holmstrom & Milgrom, 1991). Otherwise, metrics can, over time, create a "tyranny," shaping individuals and organizations in the image of the metric itself (Strathern, 2000; Muller & Muller, 2018; Choi et al., 2012, 2013). For instance, Espeland & Sauder (2007) find that when a metric gained traction among law schools, it led to persistent, reinforcing changes (in admissions requirements) to excel on the metric that came to redefine the institution and opportunity structure itself.

Yet this assessment of metrics does not fully capture the role that they can play in the innovation process. While metrics certainly can promote homogeneity and exert power as tools for assessment among R&D departments, metrics can also function much like a technology within innovation. In such contexts, metrics become part of the innovation process or 'production function,' helping to both define the task and also to assess performance in that task itself. In the next sections, I will explore this further by considering how metrics function within the process of innovation, and how such functions translate to firms with different ex-ante characteristics.

3.2.2 What Do Metrics Do in Innovation?

Innovation involves a high degree of uncertainty, and often, the challenge of choosing between multiple attractive but incompatible paths forward (Knight, 1921; Rosenberg & Nathan, 1994; Kerr et al., 2014; Gans et al., 2019). In addition to uncertainty, individuals and organizations also suffer from limits in attention (March & Simon, 1958; Simon, 1955; Cohen & Levinthal, 1990). Technological change therefore has often relied on mechanisms that can focus attention in a specific direction (Rosenberg, 1969). Metrics are one such mechanism as they can reduce extraneous information and reformulate a complex, multi-dimensional problem into one succinct target. Individuals and firms seeking to innovate can therefore align and focus R&D efforts towards a compelling and clear target, and overcome the inertia that uncertainty can create. Metrics, therefore have the potential to be inducement mechanisms for innovation (Cohen, 2010), raising the overall levels of innovative output in areas related to it.

The production of new ideas also faces a friction in terms of the costs of accessing knowledge, such as comprehension, communication and validation costs (Rosenberg, 1982; Rosenberg & Nathan, 1994; Arora & Gambardella, 1994). Metrics address this by allowing ideas to be compared against one another, creating a *lingua franca* and acting as a type of "arbiter" of scientific progress (Kuhn, 1961; Porter, 1995). Put differently, metrics are a key input in the scientific method and help quantify progress towards achieving a target.

Taken together, metrics serve a dual purpose in innovation by offering a target to align attention and resources towards, while also providing a means to quantify the advancement towards that very target. Potentially because of this latter role, metrics related to scientific and technological innovation tend to be perceived as objective, inherent features of knowledge (Porter, 1995). Yet, given that metrics are also used to set targets, they involve, at one point in time, a choice of what information to prioritize and what to lose. Rather than being facts, metrics are the product of subjective choices are made under idiosyncratic circumstances and constraints. Therefore metrics can persist with little scrutiny, even when they involve targets that are obsolete or generate negative externalities (Kuhn, 1961).

3.2.3 How Do Metrics Impact Firms' Innovation Strategies?

Shifting attention to firms, metrics are a crucial part of the R&D process that break down high-level innovation strategies into functional objectives and align resources towards them (Roth, 2009; Iansiti, 1995; Clark et al., 1987; Steen, 2017). Metrics also allow managers to evaluate different innovation approaches in a transparent and efficient manner, using progress along the metric as a decision factor (Vinokurova & Kapoor, 2020). But metrics can also have an impact on firms beyond their functional utility in R&D management within a firm.

Metrics, especially those related to scientific and technological innovation, can affect firms' broader innovation strategies when considering how they might be available to multiple firms. Whether it is due to influential firms disclosing their own metrics, or academics and regulators developing metrics, when multiple firms have access to a metric, is effectively creates a new dimension of competition upon which firms can be ranked against one another. Firms therefore face a strageic choice of whether to adopt a metric, and to what degree to focus their efforts upon it. Unpacking this question requires factoring in a firms' ex-ante domain knowledge and strategic positioning.

Firms will have different types and levels of prior knowledge (Henderson & Clark, 1990; Rosenkopf & Nerkar, 2001), which will lead them to interpret the metric different. Some with pre-existing knowledge in an area may be able to recognize that a metric is only one way to summarize a broader problem at hand (i.e. an "analogy" or "mental model") (Cohen & Levinthal, 1990; Gavetti et al., 2005; Menon & Yao, 2017), and be able to clearly see the linkages between the summary metric and the broader problem space (Helfat, 1994; Stuart & Podolny, 1996; Katila & Ahuja, 2002; Cohen & Levinthal, 1990). Firms without those ex-ante strategic investments in contrast, might interpret the metric as the main way to hone in on where to focus (Arrow, 1974; Henderson & Clark, 1990; Clark et al., 1987). Such firms risk relying on the metric as the only lens on the problem at hand and be at risk of "metric myopia."

Relatedly, firms' prior strategic commitments to the area related to a metric could also influence their response. A metric offers firms with a common lens for decomposing a problem. While this may boost overall innovation in an industry or domain by increasing knowledge flows (Mokyr, 2002, 2005), this also lowers the barriers for firms to understand and learn from each others' innovation approaches. Firms with prior strategic commitments to safety may therefore be concerned about information leakage and potential imitation. As such they may be less inclined to fully adopt the metric relative to those without such commitments.

At the same time, metrics create a tension for firms strategically oriented in a related area as they also provide a way for customers and investors to compare firms. Ranking poorly on the metric could therefore be costly to such firms. One way to reconcile this with the potential costs of imitation, is to introduce competing metrics (see Oberholzer-Gee et al. (2006) for an example). Together, this implies that firms with prior domain knowledge and strategic commitments in the area related to the metric can be expected to be relatively less inclined to adopt a new metric, compared to firm without such ex-ante investments.

3.3 Setting

To explore the impact of the introduction of an innovation metric on firms, I needed a setting that could overcome the core impediments to empirical research in the area: First, the metrics that come to be known are typically those that have successfully diffused and gained wide adoption. It is therefore, difficult to trace back to the original creation of the metric, and determine a credible set of alternative ways in the which the metric could have been defined. An ideal metric would therefore provide a bounded set of alternative areas that innovation could have developed in instead. Second, in order to study the impact of metrics on firms, it is necessary to be able to separate it from choice of the organization to adopt the metric in the first place. This required some exogenous party that introduced the metric to firms. Finally, there is limited empirical data that can link together meaningful innovation or performance outcomes to a metric over time.

This paper leverages a unique setting the addresses these empirical challenges, and allows for the estimation of the causal effect of the introduction of a metric on firms' innovation strategies: the introduction of the side impact dummy (SID) by the US government as a new metric within automotive setting. The following section provides a summary of the history of automotive safety and crash test dummies.

3.3.1 Automotive Safety

Throughout the 1960s, the US public became increasingly vocal about their discontents regarding the lack of automotive safety. [8] This led to the passage of the Department of Transportation (DOT) Act through Congress in 1967, and the creation of the National Highway Traffic Safety Administration (NHTSA) with the mandate to "save lives, prevent injuries, reduce vehicle-related crashes."[9] Today, NHTSA is the main institution charged with developing and enforcing regulations for vehicular safety in the US.

Two key NHTSA policies related to crash test dummies are the Federal Motor Vehicle Safety Standards (FMVSS) and the New Car Assessment Program (NCAP). FMVSS are *compliance* standards, setting the minimum thresholds for the design, construction, and performance of vehicles that can operate within the US. [10]

The NCAP runs tests that are based on FMVSS, but in contrast, they do not set minimum compliance thresholds, but rather *optional* targets for improvements in safety performance. NCAP tests are intended to incentivize improvements beyond the FMVSS thresholds. The NCAP also publishes the test results allowing firms, investors, and customers to assess the relative safety of vehicle models. For instance,

[8] Two figures are credited for being instrumental in mobilizing public demand for improvements in automotive safety: consumer activist Ralph Nader, who wrote *Unsafe at Any Speed: The Designed-In Dangers of the American Automobile*, and automobile entrepreneur Preston Tucker, who spearheaded the development of safety innovations citing a lack of will on the part of the "Big Three" (General Motors, Ford, and Chrysler).

[9] This was under the National Traffic and Motor Vehicle Safety Act that was enacted. NHTSA was initially named the National Highway Safety Bureau. See https://www.nhtsa.gov/ for more information.

[10] FMVSS consist of (1) a performance requirement for a given component/system, and (2) a test procedure with specific metrics and testing equipment for assessment of compliance.

whereas FMVSS 208 specifies (in section S4.1.1.3.1) that vehicles must meet safety thresholds when crashed perpendicularly against a fixed collision barrier at 30 miles per hour, the associated NCAP test requires 35 miles per hour which is a meaningful difference in velocity of impact.[11]

3.3.2 Crash Test Dummies

The first crash test subjects were not anthropomorphic test devices, but rather human volunteers and donor cadavers. Early biomechanical engineers outfitted bodies with sensors and accelerometers to measure the impact on the body of being crashed against a barrier. There were a number of limitations to this approach. For safety reasons, volunteer crashes could only be conducted at low speeds and the instrumentation had to be non-invasive. Moreover, there were issues of volunteers anticipating the crash and adjusting their body positions in ways that were unrepresentative of real-life crashes. Cadavers instead had other sorts of constraints. For one, the bodies donated for scientific research generally skewed towards elderly men with a wide heterogeneity in body types, cadaver condition, and cause of death. Because each cadaver could only be used once before degradation, the data generated was difficult to compare against one another. Furthermore, public sentiment was negative towards this use of donated cadavers.[12] Animals (including pigs and primates) have also been used in crash tests, but are not a standard approach as they face challenges in

[11] The focus of this paper is on the time period when SID was introduced by NHTSA. Though beyond the scope of this study, since 2003, a non-governmental organization, the Insurance Institute for Highway Safety (IIHS) started conducting additional crashworthiness tests based on the FMVSS/NCAP but with additional contingencies such as offset barriers and higher traveling speeds. The IIHS is a non-profit organization funded by auto insurers and insurance associations. See https://www.iihs.org/about-us for more information.

[12] See https://www.latimes.com/archives/la-xpm-1993-11-25-mn-60691-story.html for additional details. Cadaver testing does continue today, but mainly to calibrate and inform the design of new dummies and computer simulation software.

approximating human anthropometry, as well as being seated in the upright position within vehicles.

To advance the field forward, a replicable, standardized metric that could approximate humans traveling in vehicles was required. The automotive safety industry turned to the military, which had developed the first dummy "Sierra Sam" to test pilot ejection seats since the mid-1900s. At the time, the Air Force had centered on using the 50th percentile US male to develop crash dummies.[13] Following this practice, the automotive industry ended up developing their first dummy, Hybrid II for testing passive restraints in frontal crashes (seatbelts, airbags).[14][15] The development of crash test dummies in the US continued to follow this cumulative approach, when it progressed from frontal towards side impact crashworthiness.

To economize on costs and learning, SID was constructed using parts from Hybrid II by removing the arm structures and reconfiguring the thoracic cavity. Therefore, SID inherited its dimensions as a 1960s median US civilian male with a sitting height of 35" (89.9cm), a standing height of 5'8" (172 cm) and a weight of 169 lbs (76.5 kg) (AGARD, 1996; Janssen & Wismans, 1987; Searle & Haslegrave, 1969; Schneider, 1983).[16] The focus of SID was on measuring thoracic trauma as injuries to the

[13] The first-generation Sierra Sam had been engineered to model a 95th percentile US military male at the US Air Force Aerospace Medical Research Laboratory. At the time, it was a radical idea to use "a dummy that will test equipment to its maximum in all parameters" (Hertzberg, 1970), and Sierra Sam was heavily criticized for being too abstract for its intended purpose (Searle & Haslegrave, 1969). Afterwards, subsequent dummies were 'engineered to the average" of the most relevant population at the time, the average US male.

[14] Several dummies were proposed including dummies such as the VIP series, Sierra Stan, Sophisticated Sam and Dynamic Dan. These were proposed by a mixture of aircraft dummy manufacturers, medical equipment firms, government, automotive firms, universities and research institutes, and joint public-private collaborations.

[15] Hybrid II was designed by Alderson Research Laboratories and modified by GM and NHTSA for FMVSS 208.

[16] Data for the definition of the US civilian male population anthropometry came mainly from the National Center for Health Statistics survey data, under the US Department of Health, Education and Welfare's Public Health Service; and was supplemented with military data; as well as

thorax were one of the primary reasons for fatalities in side impact crashes. Injuries to other regions (e.g. neck, femur), while life-altering, were less likely to be life-ending. Thoracic trauma is composed of a measure of chest acceleration (g-force) and deflection (mm). This is calibrated using reference body sizes (height, weight) to determine thresholds for critical injury. This is because the chest acceleration and deflection that a person of a taller, broader stature can withhold before suffering injury is higher than a person of a relatively smaller stature. In summary, SID was a chosen measurement approach that yielded a measure of thoracic trauma calibrated on the 1960s median US civilian male.

3.3.3 Introduction of SID

Motivated by the significant number of fatalities arising from side impact collisions, NHTSA introduced SID as a key new measure of a vehicle's ability to protect an occupant from side impact as part of the "FMVSS 214 - Side Impact Protection Dynamic Performance Requirement" amendment.

Beginning in September 1994, all new vehicle models were required to undergo a crash test simulating a side impact collision. The vehicle would need to seat a SID in the front driver seat, and another SID in the rear passenger seat while crashing the vehicle at a speed of 30 mph (48 km/h) at a nearly right angle into a 1360 kg moving deformable barrier traveling at 15 mph (24.2 km/h) (Samaha & Elliott (2003)).

Although discussions of the need for improved side impact protection had been documented from at least the early 1980s, the ability for NHTSA to introduce new testing requirements was contingent on Congressional budget approval. Thus, the exact timing of when NHTSA would have the resources to implement SID in a

experimental data collected by Alderson Research Laboratories on volunteer subjects between ages 16-70 for variables otherwise unavailable. See Starkey et al. (1969) for full details.

FMVSS was difficult to anticipate, even for NHTSA Directors themselves (NHTSA (1995)). NHTSA also communicated its intent to launch a concurrent side impact test using SID as in the FMVSS with a higher impact speed of 35 mph. This was publicly available information so firms were aware that SID was going to be used not only to assess the minimum threshold of side impact protection, but also the relative performance.[17]

To supplement the data collected from reports and studies relating to automotive safety and crash test dummy development, I also interviewed program administrators, researchers contracted by regulators, senior engineering leaders at automotive firms, and consumer advocates who provided additional context. Two key aspects to highlight are that first, the US DOT was a leader at the time among global regulators for automotive safety. Therefore, even firms headquartered abroad or with mainly non-US revenue streams were still responsive to SID:

> In the US, regulators are fairly transparent...The US led the way with crash dummies and passing regulations. This was very effective and flushed out all the problems with the dummies, then Europe, Japan, China would follow... They would adopt the same dummies and by then, they would be very well vetted.
>
> A dummy is designed to gauge how safe a vehicle is based on the [US] FMVSS ...Tons of money and time go into getting a dummy ready...Other countries like China use the same dummies and just tweak thresholds to their speed limits.

[17] Congress however, rejected its allowance when proposed for the budget in 1994 and 1995, so it was SID did not debut in the NCAP until September 1996. Taken together, vehicles entering the market after September 1994, or in other words, with production years 1995 onwards were affected by the new measurements captured by SID.

A second aspect of the setting the interviews surfaced was that SID was not a compliance issue for firms. While SID was incorporated in the mandatory FMVSS, it was intentionally designed to be met by all the vehicles in the market. The incorporation of SID into the other NHTSA program, NCAP, instead, was designed to be at a level that could generate a discernable difference between vehicles in safety performance. Therefore the regulatory goal was to incentivize improvements in safety by ranking firms against one another rather than through compliance. Firms therefore had the autonomy to choose whether or not, and to what degree, to focus their innovative efforts in improving along the metric:

> From NHTSA's perspective, it was worthwhile to set the level of [FMVSS] regulations lower. This gave newer cars a chance to get in the market as they just needed to get through the minimum regulation thresholds ... essentially all vehicles meet the FMVSS.

3.3.4 Research Design

The use of SID as a new metric for side-impact crashworthiness protection in the US automotive industry provides two sources of variation that allow for estimation of the causal effect of introducing a metric on firm safety performance and innovation outcomes.

First, when SID was introduced, not every vehicle model could adopt SID into its design immediately. Instead, there was significant variation in the stage of the product life cycle that each vehicle model was in. At a basic level, a vehicle model follows three phases in its product life cycle: (a) it is introduced as an all-new design, (b) it undergoes several mid-cycle refreshes with minor styling updates, and (c) it undergoes a major redesign (i.e. a "generation") and eventually is retired.

Automotive firms (commonly referred to as “makes”) usually have multiple vehicle models in the market at a time, but plan for them such that not all of the models are due for a redesign at once. This allows them to spread out risk and avoid straining attention and resources. On average, a model will stay in production with only minor refreshes for 4-5 years before undergoing a major redesign (Hill et al. (2007)). Therefore, at the time that SID was introduced, there were some vehicles in the process of being redesigned or introduced (herein, referred together as redesigned) that could respond more quickly, and others that had recently completed a redesign and were not able to adopt SID into their designs until later.

Second, in addition to the quasi-experimental variation in the timing of adoption of SID by vehicles, the study also leverages differences in the body type of occupants in the driver and rear left passenger seat from that of SID. While SID was designed based on the 50th percentile civilian male in the 1960s – the most common vehicular occupant on US roads at the time – by the time SID was introduced, the demographics of drivers and passengers on the road had become much more diverse: in 1994, men only represented 50.85% of licensed drivers.[18] Moreover, even the average US civilian male was undergoing changes during that time, with a roughly 10% increase in weight to 182.4 lbs (82.9 kg) and 1.5% increase in height to 5’9" (175.8 cm).[19] Therefore, at the time of introduction, there was natural variation in how representative SID was for a given occupant in a driver or front rear passenger seat. This allows for the empirical exploration of whether the metric had a focusing effect in terms of safety performance improvements. Furthermore, the setting is ideal in that there is rich data that can link together the metric with innovative and performance

[18] This percentage dipped below half in states such as Alabama, Delaware, Georgia, and Michigan among others. See https://www.fhwa.dot.gov/ohim/1994/section3/dl1a.pdf

[19] https://www.cdc.gov/nchs/data/ad/ad347.pdf

outcomes as described in the next section.

3.4 Data and Empirical Strategy

Broadly, I needed several types of data to explore the relationship between SID as a new metric for safety, and the subsequent firm responses. First, I needed to observe the safety performance outcomes of vehicles in side impact accidents before and after the introduction of SID. Second, relatedly, I needed to collect covariates for such accidents to control for omitted variable bias. Third, to quantify the timing variation in the incorporation of SID into a vehicle model, I needed to collect data on major redesigns by model and production year. Fourth, to assess variation in the occupants' distance from SID in terms of body type, I was interested in collecting occupants' height and weight data. Finally, to understand differences in the impact of SID on firms, it was also essential to collect measures of the level of strategic investments and domain knowledge in safety ex-ante of SID across different firms. Note that in this study, firms are defined as makes. This is because although makes are typically owned by larger parent companies, in practice, parent company ownership varies over time with makes typically setting and following their own strategies. For example, Toyota, Lexus and Daihatsu are separate makes, that are controlled by the same parent company, Toyota Motor Corporation. This section below describes the primary sources of data in detail and summarized in Table 3.11. Further information on firm safety investment variables are discussed in Section 3.5.4.

3.4.1 Crashes

The main data source for this study is the NHTSA National Automotive Sampling System (NASS) - Crashworthiness Data System (CDS) data set for years 1991-2007. The CDS data is compiled such that the primary unit of analysis is the accident-vehicle-occupant level. Two features of the CDS make it ideal for understanding the impact of SID. First, the CDS is a representative and random sample of all police-reported vehicle crashes in the US involving property or personal damage (i.e. a hazardous event). Second, unique to the CDS, there are field research teams based at 24 geographic sites that collect data from police reports, on-sight expert investigations, and occupant interviews. This produces detailed information on vehicle characteristics, accident conditions, and occupants including demographics and driving impairments. Importantly, because of the occupant interviews, the CDS includes data on the height and weight for all occupants involved.

The full dataset is restricted in three ways to hone in on the main sample affected by SID. First, the sample is limited to accidents involving passenger cars, dropping trucks, motorcycles, recreational vehicles (e.g. ATVs, mopeds, dirt bikes), buses and unknown vehicles (e.g. hit and run). Second, the sample is limited to those involved side-impact crash accidents. Third, it is limited to occupants in the front driver and left rear passenger seats as these are the occupant positions where the FMVSS and NCAP required SID to be seated during testing.

Alternative Crash Data

Two additional data sets on vehicular accidents in the US are used to assess the robustness of the main findings. Additional details about the alternative data sets, and how the trade-offs they involve relative to the main CDS data set are summarized in Table 3.23.

The first data set is the Fatality Accident Reporting System (FARS) database which is the primary source used in prior social sciences research (Levitt & Porter, 2001a,b; Grabowski & Morrisey, 2004). The advantage of this data is that it captures the full population of vehicular accidents in the US that involve a fatality. At the same time, there are concerns about sample selection as vehicles involved in potentially fatal accidents that do not result in a death are excluded. To mitigate this issue in part, I employ the strategy proposed in Levitt & Porter (2001b) when using the FARS data for testing robustness by restricting the sample to occupants in vehicles involved in accidents where there is a death in the other vehicle. This approach results in a subsample where incurring a fatality is not the sole selection criterion. Another limitation is that the height and weight are only collected for the drivers, leaving out rear left-side passengers that were also affected by SID.

The second data set is the General Estimates System (GES) database. Similar to the CDS, it is a representative sample of police-reported motor vehicle crashes across the US, however it is larger in quantity of observations. This is because it only draws information from police reports without additional follow-up or on-site data collection. As the name implies, the intention is to estimate the number of accidents of different types. A key limitation for the purposes of this study is that the heights and weights of occupants are not collected. To ameliorate this, data from the US CDC Behavioral Risk Factor Surveillance System (BRFSS) was used to impute occupant height and weight. The BRFSS collects survey data to get a representative sample of US residents and their health behaviors and conditions.[20] The imputation was conducted using the occupant location, age, gender, year, and whether their zip code corresponds to a low-income area.

[20] https://www.cdc.gov/brfss/index.html

3.4.2 Vehicles Specifications and Redesign Timing

To collect data on major redesign timing for vehicle models, I used the Wards Automotive Yearbooks. Wards, now owned by Informa, publishes industry trade information and magazines covering the US automotive industry since the mid-1920s. The Wards Automotive Yearbooks contain summaries of every new and majorly redesigned vehicle models launched in the US (as well as Canada and Mexico). This was non-trivial because for the time period of study, the Yearbooks are only available in hardcopy, and the information was contained in free-form, non-standard text. As such this data on new vehicle model launches and redesign timing were hand collected. To verify and supplement this data, web searches were conducted for each vehicle make and model appearing in the CDS using Wikipedia, Edmund's and KellyBlueBook.

The Wards Automotive Yearbook also contains detailed information on vehicle specifications and the corresponding manufacturer suggested retail price. For some of the vehicles in the CDS and FARS datasets, the vehicle identification number (VIN) was available. These were run through the NHTSA Product Information Catalog and Vehicle Listing (vPIC) VIN Decoded API to collect basic information (e.g. curb weight, number of doors). This is a sparse manufacturer-provided dataset, but where available, it was used to validate the specifications data from WARD's Automotive Yearbooks.

3.4.3 Summary Statistics

Table 3.1 presents descriptive statistics of the sample: The average accident occurred in June-July 2000 in an area with a speed limit of 51 miles per hour under dry, light daytime conditions with no alcohol or drug involvement. The average occupant was a 39 year old male with a height of 170.4 cm and 76.3 kg. The average vehicle was worth \$25,073 at the time of purchase and had a production year of 1993.

Taken together, the main sample for analysis is composed of 22,715 accident-vehicle-occupant observations that are randomly and representatively drawn from police-reported motor vehicle side impact crashes occurring in the US from 1991 to 2007. 16,573 unique occupant types (age, height, weight, gender,race) and 3,308 unique vehicles (make, model, model year) were involved in these accidents. The vehicles involved in these accidents are produced by 46 unique firms.

3.4.4 Estimating Equations and Identifying Assumptions

The main estimating equation uses a difference-in-differences specification:

$$\begin{aligned} E[\text{Fatal}_i|X_{ik}, \text{PostMetric}_{jt}] = \ & \beta_0 + \beta_1 \text{PostMetric}_{jt} + \beta_2 X_{ik} \\ & + \delta_j + \gamma_t + \eta_{m(j),t} + \epsilon_{ijtk} \end{aligned} \tag{3.1}$$

where Fatal_i equals 1 when occupant i traveling in vehicle model j produced in year $t(j)$ involved in accident $k(i,j)$ dies (expressed as t and k respectively for clarity of notation), and 0 otherwise. PostMetric_{jt} equals 1 when vehicle model j undergoes its first major redesign after the introduction of SID in year t and thereafter, and the key coefficient of interest is β_1. The controls include X_{ik}, a vector of accident and occupant controls described in Table 3.1; and vehicle model fixed effects (δ_j), production year fixed effects (γ_t), and vehicle make and production year fixed effects ($\eta_{m(j),t}$).

In this specification, I am comparing the difference in likelihood of fatality when getting into a side impact accident in a vehicle model that has incorporated SID into a major redesign, with vehicle models that have not yet incorporated SID into its design. If vehicle models that incorporate SID have a lower (higher) fatality rate, then the estimate of β_1 will be negative (positive). The vehicle model fixed effects included in this specification account for the non-random sorting of vehicles into accident conditions. The vehicle make and production year fixed effects account for any shocks or changes among makes in certain production years. The production year fixed effects absorb variation in the costs of incorporating safety innovation and design over time.

To understand how the relationship between safety outcomes and SID varies depending on its distance from the occupant body type, the study estimates:

$$\begin{aligned} E[\text{Fatal}_i|X_{ik}, \text{PostMetric}_{jt}] = \; & \beta_0 + \beta_1 \text{PostMetric}_{jt} + \beta_2 \text{SIDOcc}_i \\ & + \beta_3 \text{PostMetric}_{jt} \times \text{SIDOcc}_i \\ & + \beta_4 X_{ik} + \delta_j + \gamma_t + \eta_{m(j),t} + \epsilon_{ijtk} \end{aligned} \tag{3.2}$$

where SIDOcc_i is an indicator equal to 1 when occupant i is close in height and weight to SID; and the other variables are defined as above in Equation 3.1.

For both models, a key identifying assumption is that conditional on the included controls, changes in the independent variable (PostMetric_{jt} and SIDOcc_i) are not correlated with unobserved determinants of safety and that the composition of treated and control groups are stable. Exploration of the data supports these assumptions: While unlikely given the industry cost structure of a vehicle redesign, one potential concern could be that firms shorten their redesign time to respond to SID. Figures 3-2 (a) and (b) reduce this concern as both the average vehicle model redesign time and the average time between a new vehicle launch or major redesign being launched by a firm are stable, and in fact, slightly increase after SID. Similarly, another concern could be if the demographics of drivers shift such that the number of SID-like vehicle occupants is inconsistent. Figure 3-3 (a) and (b) show that the distribution of height and weight of occupants in the driver and rear passenger seat have no significant differences before and after SID, though the mean weight is slightly lower in the post-period. Finally, as presented in Table 3.27, fatality rates from frontal and rear-end crashes see no statistically significant change related to a major redesign following SID suggesting that there is no confounding, omitted variable that is increasing safety at the same time.

3.5 Results

3.5.1 Did the Metric Increase Performance?

Using only the raw data, there is a correlation between vehicle models incorporating the metric SID into their design (i.e. either a new model or having had a major redesign after the introduction of SID), and a decline in the number of side fatalities. This is presented in Figure 3.6 which is a binned scatterplot where the mean occupant fatality is plotted against 30 equal-sized bins of vehicle production year.

Table 3.2 presents the estimates from formalizing this in a regression (Equation 3.1). Column 1 is a baseline regression including only a dummy that equals 1 when a vehicle undergoes its first major redesign after the introduction of SID, as well as vehicle model, and production year fixed effects. The following columns progressively adds controls listed in Table 3.1. Column 2 adds occupant characteristics as controls. Column 3 includes vehicle controls. Column 4 adds accident controls and a dummy for each accident year.

In Column 5, post-double-selection LASSO is used to augment the regressions in Columns 1 to 4 by using the LASSO method to optimally select regression control variables. PDS-Lasso helps mitigate potential bias in the selection of control variables, both in terms of overfitting, omitted variable bias and potential author-introduced bias in the selection of controls, given the high-dimensional accident data (Belloni et al., 2014). This involves three steps where first, LASSO is used to estimate and select controls for:

$$\text{Fatal}_i|X_{ik} = \beta_1 x_{i,1} + \beta_2 x_{i,2} + ... + \beta_p x_{i,p} + \epsilon_{ijtk}$$

Where $x_{i,1}, \ldots, x_{i,p}$ are all potential control variables (excluding the independent variable of interest, PostMetric$_{jt}$). In this analysis, I used 102 variables including all of those listed in Table 3.1, the squares of those variables, as well as additional controls such as the number of vehicles involved in the accident, number of occupants in a vehicle, whether the occupant is pregnant, the day of the week of the accident, speed limit, occupant race dummies, weather condition dummies, impact angle, vehicle body type, other vehicle's curb weight, number of lanes, etc. Second, LASSO was used to estimate and select controls for:

$$\text{PostMetric}_{jt} = \beta_1 x_{i,1} + \beta_2 x_{i,2} + ... + \beta_p x_{i,p} + \epsilon_{ijtk}$$

Then finally, the superset of controls selected in the first two steps (i.e. had non-zero coefficients for both regressions) used in a linear regression with Fatal$_i|X_{ik}$ as the outcome variable, and PostMetric$_{jt}$ is the focal independent variable for causal inference.

The resulting controls selected by PDS-Lasso, presented in Table 3.13, were occupant age squared, occupant height and height squared, the number of vehicles involved, whether the occupant was wearing a seat belt, whether there was alcohol involvement, and whether there was drug involvement. Vehicle model, production year, and accident year fixed effects were included as unpenalized regressors throughout.

Overall, the estimates are quite similar in magnitude across all specifications with standard errors clustered at the make level. Focusing on Column (4) with all accident, vehicle and occupant controls; and Column (5) with PDS-Lasso controls shows a sharp and significant decrease in side impact fatalities by 48-50% after the introduction of SID. This estimate is similar in magnitude to the estimated improvements from seat belts (45-60%) [21][22] along with airbags (61%) [23].

To test for violation of parallel trends, in Figure 3-5 I plotted the estimates from an event study specification in which occupant fatality is regressed onto 10 interaction terms between the treatment status and the number of years before and after the first major redesign after the introduction of SID. Visual inspection finds that the pre-treatment trends are not statistically significant from zero. See Appendix 3.6 for additional analyses of the pre-trends using Roth (2020).

The relationship between the introduction of SID and the reduction in side-impact fatalities is robust to using a probit regression specification as presented in Table 3.14. The results are also present when analyzing the alternative accident data sets of FARS and GES as presented in Table 3.25.

Taken together, the introduction of SID as an innovation metric within automotive safety led to an improvement in in vehicles' safety performance in the direction of the measure on average. This is in line with the conceptual framework in that introducing an innovation metric provides firms with a target to align resources around, and the ability to quantify progress and compare actions against one another. At the

[21] https://www.cdc.gov/transportationsafety/seatbeltbrief/index.html
[22] https://injuryfacts.nsc.org/motor-vehicle/occupant-protection/seat-belts/
[23] https://www.iihs.org/topics/airbags

same time, given the loss of information and context that comes with this process, it raises the question of whether occupants being close versus far in body type from SID had any effect on their likelihood of incurring a fatality in a side-impact crash.

3.5.2 Did the Metric Focus Effort?

To probe whether safety performance outcomes vary for occupants whose body types more closely match that of SID, I estimate Equation 3.2 which separates out those occupants that are relatively close to SID from those more distant.

Table 3.3 presents the results with Columns (1) through (4) following a progressive addition of the controls in Table 3.1, and Column (5) including PDS-Lasso selected control variables as in Table 3.2. In the main analysis, SID-like occupants are defined as a dummy which is equal to 1 when the occupant is within 0.5 standard deviations in height and weight from SID. Columns (4) and (5) find that SID-like occupants (who start from higher baseline fatality rates) benefit from an disproportionate, additional decrease in fatality rates relative to those for whom SID is less precise.

This result is robust to different definitions of SID-like occupants as shown in Table 3.18 which tests the sensitivity of the results to the definition of SID-like occupants by relaxing and restricting the thresholds of height and weight (e.g. 0.6 × 1 standard deviation in height and weight, etc.). The results also hold using a probit specification in Table 3.15, and two alternative datasets in Table 3.26.

Moreover, the results in Table 3.3 do not hinge on gender. Table 3.16 splits the sample into males and females, finding that the additional improvement in safety outcomes for SID-like occupants is present across both genders. The number of female SID-like occupants is however, lower than male SID-like occupants as females in general tend to be smaller in stature than males. This reconciles this with recent studies (Perez, 2019; Linder & Svensson, 2019) documenting that females tend to have higher fatality rates than males: metrics and the focusing effect they can create (and loss of information) in the upstream innovation process can lead to gender (and other sources of minority group) bias in downstream markets.

In addition, I estimate a model that interacts the treatment dummy with all the occupant, vehicle and accident controls. I employ the sorted partial effects method (SPE) developed by Chernozhukov et al. (2018) to visualize and conduct a classification analysis of the partial effects. This allows me to explore heterogeneity in the partial effects beyond the average treatment effect. Details and the full analysis are in Appendix Section 3.6. I find, using the SPE approach, that there is a wide heterogeneity in the returns to SID: some occupants benefit significantly more than others, whereas others are worse off. Looking at the characteristics of the group with the largest gains from the introduction of the metric, I find that the median height and weight track closely to that of SID. This suggests that even with relaxing the assumption of linearity and an exploration of the partial effects, there is evidence that occupants close in body type to SID gained a disproportionate reduction in fatality rates. The introduction of an innovation metric led firms to target efforts towards it, but perhaps because of a loss of information and context, resulted in a focusing effect around the narrow definition of the metric.

3.5.3 Did These Improvements Require Meaningful Changes to R&D?

One question these findings may raise is whether the improvements in safety performance required firms to innovate in meaningful ways in their vehicle design and development. To test this, I employed car specifications data to estimate the impact of SID on the physical structure of vehicles. I find that firms had to make meaningful changes to the structure of their vehicles to improve side crash protection.

Table 3.5 presents estimates using OLS regressions of the following estimating equation:

$$\begin{aligned}\text{SPEC}_{jt} = \; & \beta_0 + \beta_1 \text{Post}_t + \beta_2 \text{Redesign}_j \\ & + \beta_3 \text{Post}_t \times \text{Redesign}_j + \beta_4 X_j + \epsilon_{jt}\end{aligned} \tag{3.3}$$

where SPEC_{jt} corresponds to the car specification (e.g. vehicle height) for vehicle model j produced in year t; Post_t equals 1 when the production year of vehicle j is after the introduction of SID; Redesign_j equals 1 when the vehicle model j has undergone a major redesign or is newly launched; X_j is a vector of vehicle controls including model fixed effects; and ϵ_{jt} is an error term.

Column (1) shows that when a vehicle undergoes a major redesign after the introduction of SID, it increases slightly in width by 0.3%. Similarly, Column (2) shows that there was an increase in weight by 1.3%. Columns (3) and (4) finds no change in the overall height and wheelbase of the vehicle. Column (5) finds a slight 0.4% increase in length which may reflect design changes in tandem with the adjustments to width and weight. Column (6) finds no change in the horsepower of the vehicle in response to SID. Column (7) finds that the miles per gallon of the vehicles actually decreased by 1.2% in major redesigns occurring after the introduction of SID. This points to a potential costly design trade-off to prioritize safety through increasing the size of the vehicle structure to minimize intrusion, and curb weight through the use of more crash-resistant materials, at the cost of fuel economy. All regressions in Table 3.5 control for the retail price, number of doors, as well as vehicle type (hatchback, convertible) and vehicle model fixed effects.

This is consistent with qualitative interviews with industry experts who described how firms had to rethink the door padding and materials, structural stiffness, and passenger compartment size in response to SID to protect occupants. This corresponded in many cases to an increase in width, length and weight of the vehicles:[24]

> If you are looking at way to transport people safely it's exactly the same way you ship fragile china... You must start with a strong box... in early side crash tests, structural collapse was very common... There were too many weak boxes.
>
> [Post-SID], the focus was on making sure compartments stay intact... There was a huge difference in car performance on intrusion than before, and the development of crumple zones.

Table 3.6 validates the qualitative interview suggestions, finding that increases in width, weight and length reduced fatalities in side impact crashes, while improvements in fuel economy (miles/gallon) are associated with worse fatality rates.

Taken together, these findings suggest that the new metric, SID, required firms to make changes to their vehicles, including some that are significant enough to observe in the overall vehicle structure. Moreover, these changes in width, weight and length associated with the introduction of SID came at the cost of worse fuel economy, pointing to the metric necessitating meaningful innovation and product design trade-offs.

3.5.4 How Did Firms Respond Differently to SID?

The results thus far have found that *in aggregate*, the introduction of an innovation metric in automotive safety led to improvements in side impact protection, as well

[24] Figure 3-8 illustrates the structural changes in the Mercury Grand Marquis before and after the introduction of SID.

as an additional, disproportionate benefit for occupants that were more close to the narrow definition of the metric SID itself. Furthermore, these performance outcomes were the result of significant innovation and changes.

Firms however, as highlighted in Section 3.2, are expected to respond differently to metrics depending on their prior strategic commitments and domain knowledge related to the metric. This is particularly applicable to the automotive industry where firms have historically prioritized saliently different strategic areas: for instance, some firms emphasize cutting-edge design, some firms focus on affordability, while others prioritize safety as a source of strategic advantage.

To explore this empirically, I needed to separate out leading firms that had made strategic commitments and possessed domain knowledge in safety prior to the introduction of the metric, from those that had not. To understand firms' strategic positioning in safety, I collected data on management's intended strategic positioning (shareholders reports), and the publicly perceived strategic positioning (news). To understand firms' domain knowledge, I collected data on their internal R&D capabilities (patents) and product engineering and production capabilities (defect investigations).

Firms' Safety Positioning

To understand a firm's external positioning in regards to safety, I collected data on how it positions itself to its current and potential shareholders, and how it is perceived by the media. I used the Mergent Archives data and Dow Jones Factiva database as sources for measuring how a firm positions itself and its perceived positioning, respectively.

Annual shareholders reports data for the three years preceding the introduction of SID were collected from Mergent Archives for each parent firm. Mergent Archives is a database of scanned annual reports with wide coverage of corporate annual reports from 1909. Once digitized, I extracted the text from the reports excluding the text related to the financial statements and auditor reports. Then I used text classification techniques described in Appendix Section 3.6 to determine the fraction of words in the shareholders report that are safety-related language.

Media coverage on each firm (make) was collected from the Dow Jones Factiva platform. Because the objective was to capture the publicly perceived strategic positioning of a firm, I focused on new vehicle launches using the Factiva filter of “New Products/Services” and “New Product Approvals” rather than all articles (that could include policy or broader industry issues, etc.). All such articles published in a newspaper outlet in the US in 1990 to 1993 with a vice-president or president/CEO mentioned or quoted were included. The text was then classified to determine the fraction of words across all the articles that are safety-related. These data were then used to develop measures of whether a firm was above the median versus below the median in terms of ex-ante strategic commitments to safety.

Firms’ Safety Knowledge

To understand a firm’s prior knowledge in safety, I collected data on its innovation outcomes in safety, and its product safety performance outcomes. I used the US Patents and Trademark Office PatentsView data and the Office of Defects Investigations data as paper trails of firm’s innovation and product development capabilities in safety, respectively.

The PatentsView database is a joint initiative from the USPTO providing data on all patents applications and grants in the US. To build a sample I needed to link each firm to its patent assignee ids. This was aggregated at the parent level for consistency (e.g. Toyota Motor Company rather than Lexus). To do so, I first collected all the patent assignees of patents under CPC code B60 (Vehicles, Vehicle Fittings, Vehicle Parts). For firms publicly listed in the US, I linked the assignees to firms using the Compustat bridge developed by Autor et al. (2020a). For other firms, I used fuzzy name matching between the assignee name and the firm name. To ensure I captured all the permutations of a firm's name I collected all the subsidiary names for each firm using the NHTSA vPIC API.[25] For instance, using vPIC I found that Honda Motors Company also did business as Honda of America Manufacturing, Honda Manufacturing of Alabama, Honda of Canada Manufacturing, Honda Manufacturing of Indiana, etc. Then the bridge between firms and assignees was used to collect the universe of patents granted to each automotive firm. To assess the safety orientation of each patent I extracted the patent summary text, which includes the title and a condensed description of the invention, and used text classification techniques to determine if a patent is safety-related or not. Appendix Section 3.6 provides a detailed description on this method.

While patents represent safety capabilities in upstream innovation, firms must also invest in translating such innovations into safe products. To measure this engineering and production capability, I needed a dataset that could capture the safety quality of vehicles produced by the firms. I drew upon a comprehensive defects investigations dataset from the DOT Office of Defect Investigations database. A defect, as defined by the ODI, is any issue "in performance, construction, a component, or material of a motor vehicle or motor vehicle equipment" that affects the ability of a vehicle to "protect against unreasonable risk of death or injury in an accident." These defect investigations are precursors to recalls (which can involve strategic decisions

[25] https://vpic.nhtsa.dot.gov/

from regulators and firms), and are initiated after a sufficient number of consumer complaints and petitions are made to NHTSA alleging a safety defect; and the ODI has made a technical analysis that corroborates the need for further information.[26] For each firm, I use the ODI to collect the number of vehicle models that have ever had a defect investigation opened up about it; and took the ratio of defective vehicle models to total vehicle models.

Firm-Level Results

To unpack how firms with strategic commitments in safety responded to the introduction of SID relative to firms without I estimated Equation 3.2 splitting the sample into firms with below versus above the median investments in positioning themselves as leaders in safety. These results are presented in Tables 3.7. Columns 1 and 2 look at firms with lower orientation around safety versus those with higher as measured by the ratio of their annual reports text that contained safety-related language. Columns 3 and 4 splits the sample into firms with a lower versus higher ratio of safety-related language in the media news coverage of their new product launches. Given the level of measurement, standard errors are clustered at the parent level for shareholders reports, and clustered at the make level for media coverage. Firms that position themselves as less safety-focused in their text to shareholders, as well as those perceived by the media as placing less emphasis on safety, mainly improved in safety performance within the narrow definition of the metric.

Shifting to the prior knowledge of the firm, Table 3.8 presents estimates splitting the sample into firms with below versus above the median prior knowledge in the space of automotive safety. Columns 1 and 2 explore how firms with relatively

[26] https://www-odi.nhtsa.dot.gov/recalls/recallprocess.cfm

lower safety patents compare to those with higher. In Column 1 I find that those with lower innovation capabilities in safety focus primarily on the metric, and make improvements that largely reduce fatality rates for SID-like occupants. In Column 2, in contrast, those with higher safety innovation capabilities are less focused on the metric, and make improvements that benefit all types of occupants beyond just SID-like ones. Columns 3 and 4 splits firms into those that have relatively lower product engineering capabilities in terms of safety versus higher. Column 3 finds that firms with more defect investigations, consistent with those with lower patents, focus their improvements on the metric SID. Column 4 finds that those with fewer defect investigations are less inclined to focus solely on SID in their products.

Taken together, Tables 3.7 and 3.8 suggest that there is a dichotomy in the responses that firms have to the introduction of an innovation metric. Whereas some firms make changes that lead to improvements in the *broad direction* of the metric, leading to lower fatality rates for occupants of all body types, others concentrated their efforts – perhaps overly so – only on the metric itself. This response seems to be moderated by the degree of strategic positioning and knowledge in safety a firm had invested in prior to the introduction of the metric.

Table 3.9 extends the analysis by examining the intersection of strategic positioning and knowledge. The first two columns look at firms with below the median safety positioning in their annual reports, while latter two examine above the median firms. The sample is further split into those with below (Columns 1 and 3) versus above the median safety-related patents (Columns 2 and 4). I find that firms with low knowledge are driving the significant focusing effect on SID. This suggests that having prior knowledge is the key mechanism that mitigates metric myopia. This may be because given the loss of information inherent in developing a metric, SID can not fully capture the nuances of the problem of side impact safety. Firms that use SID without prior knowledge in safety may find that while it can be a useful target to accelerate innovation and quantify progress with, they are more susceptible to overfocusing on the narrow definition of the metric itself.

3.6 Conclusion and Future Research

> "*Man is the measure of all things: of those that are, that they are; and of those that are not, that they are not.*" (Protagoras, 490–420 BCE)

This insight, traced back to at least Protagoras' time, surfaces that all metrics are born from some individuals' choices of what to measure and how to define it. Over time however, as metrics become more widely adopted, the initial subjective choices can be abstracted away. This leaves metrics to be perceived as objective arbiters of knowledge. Given that metrics are ubiquitous within firm R&D and science, and can sometimes persist for decades or even centuries, it is important to understand how they shape innovation. This paper provides some of the first empirical evidence examining the role that the introduction of a metric can have in shaping firm innovation.

I start by proposing a conceptual framework to understand how metrics might impact the process of innovation. By reducing dimensionality and providing consensus on what constitutes meaningful progress, a metric allows firms to overcome uncertainty when pursuing innovation. Firms can align their resources towards a common target, and accelerate their innovative output. At the same time, the very simplification that make metrics useful for innovation can also raise a potential cost. Because the process of using a metric in and of itself involves a loss of information, metrics that become reified as fact can lead some firms towards an overly narrow focus and innovation trajectory. This can lead to heterogeneous effects for firms depending on their ex-ante characteristics. Firms that possess knowledge or have made strategic commitments to the area related to the metric are less likely to suffer from "metric myopia," instead using the metric to accelerate overall innovation while still maintaining their own thesis on how to improve the problem. Firms without such prior investments are more at risk of relying too much on the metric to define the problem, which could lead them to pursuing an overly narrow trajectory of innovation.

To explore this empirically, I use rich data on vehicle accidents in the US combined with hand-collected data on car specifications to estimate the impact of the introduction of a new metric (SID) on firm innovation in automotive safety. I exploit quasi-experimental variation in the timing of adoption in vehicle models, and differences in how representative the metric was for different occupants to establish a causal estimate of the impact on fatality risk in aggregate and across different occupants. Results show that SID led to significant improvements in protection against fatalities, but also to disproportionate benefits for occupants similar to SID. Furthermore, these changes were driven by heterogeneous innovative responses from firms. Consistent with the framework, firms that made prior strategic investments and domain knowledge in safety R&D made improvements for all occupants. In contrast, firms that did not possess this prior experience with the problem were focused solely on making improvements for SID-like occupants and made costly trade-offs in their vehicle R&D to achieve this.

Taken together, this paper offers three main contributions. First, it documents the empirical relationship between the introduction of a metric and firm innovation outcomes in a quasi-experimental setting that allows for a causal interpretation. This setting is economically significant and relevant for policy and social welfare. Second, this paper highlights that while metrics can induce innovation output to increase and can also shape the direction of innovation effort, these average effects do not translate to every firm. In fact, this study highlights that metrics can have quite a different impact across different firms. Some firms are better placed to leverage the benefits of a metric while others are at a disadvantage based on their prior characteristics. This emphasizes the impact that metrics —even those that may appear to predominantly exist within scientific or technological domains —can have for firms' broader strategy and management decisions. Third, while previous work has focused on metrics as part of standards, the paper highlights that metrics involve their own, unique mechanisms for innovation, and are an important area of focus for future work.

More broadly, this paper also offers implications for diversity and inclusion. Bias in product design has been documented in a number of settings, ranging from consumer goods to safety to health care (Goldberg, 2002; Perez, 2019; Liu & Mager, 2016; Linder & Svensson, 2019). This paper highlights metrics as a potential source of this bias. Metrics chosen out of convenience or idiosyncratic conditions may end up shaping innovation, even without any intention of bias, in a way that excludes people more distant from the metric definition (e.g. women, visible minorities, elderly). Furthermore, because such metrics are situated upstream in the R&D process, they may be commonly overlooked and difficult to trace back to. Managers and policymakers seeking to use metrics to shape innovation should consider mitigating this risk by introducing multiple metrics at once, launching metrics in tandem with resources on the broader problem to be solved, and fostering a culture of iteration in the metrics that are used.

Future research can also explore how metrics shape innovation depending on the stage of maturity. The findings in this paper imply that in the earlier stages of a problem space when innovation involves uncertainty, a metric is a *complement* to problem understanding. Potentially, as an area matures and a shared understanding of the problem is developed, a metric may be able to substitute for deeper domain expertise. In such cases, a metric would no longer be a dimension of competitive advantage but rather a standard prerequisite for participation in that area. This suggests a potential dynamic interplay between metrics and problem understanding in firms. Another area for future work is to examine settings when there are numerous metrics available to firms. While this setting focused on the introduction of one metric, the growing proliferation of software-driven R&D and lower costs of collecting "big data" raises questions of how multiple metrics might influence innovation, whether there is an optimal number to introduce, and what the benefits versus costs are of providing more frequent and higher precision metrics in the innovation process.

In summary, this research suggests that metrics in science and technology, although not typically considered part of strategy formulation, are powerful mechanisms that can shape the rate and direction of innovation for individuals, firms, and industries, sometimes for centuries. This paper provides empirical evidence in line with this, and highlights this as an exciting area for future work to further our understanding of this overlooked, important lever for innovation.

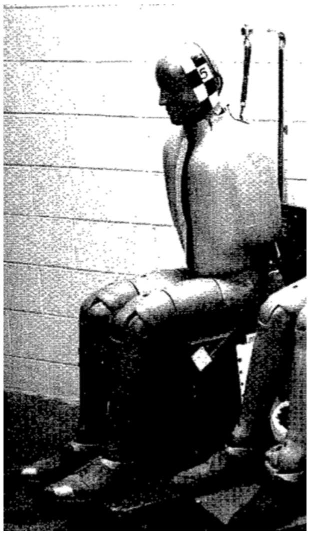
Tables and Figures

Table 3.1: Descriptive statistics for the main crash sample

Variable	**Mean**	**Std. Dev.**	**Min.**	**Max.**
Accident Controls				
Accident Year	2000	4.978	1991	2007
Accident Month	6.797	3.434	1	12
Speed Limit	51.659	10.461	15	80
Wet Conditions	0.148	0.355	0	1
Dark Conditions	0.233	0.422	0	1
Daytime	0.811	0.391	0	1
Alcohol Involved	0.109	0.311	0	1
Drugs Involved	0.043	0.203	0	1
Occupant Controls				
Age	39.255	18.111	18	97
Male	0.503	0.5	0	1
Height (cm)	170.416	10.168	119	220
Weight (kg)	76.291	17.855	36	136.078
BMI	26.264	5.871	15.036	69.822
Pregnant	0.014	0.118	0	1
No Seat Belt	0.151	0.358	0	1
Vehicle Controls				
MSRP	25073.03	13018.78	6907.5	129697.5
Car Age	6.546	4.949	0	22
Model year	1993	6.765	1970	2008

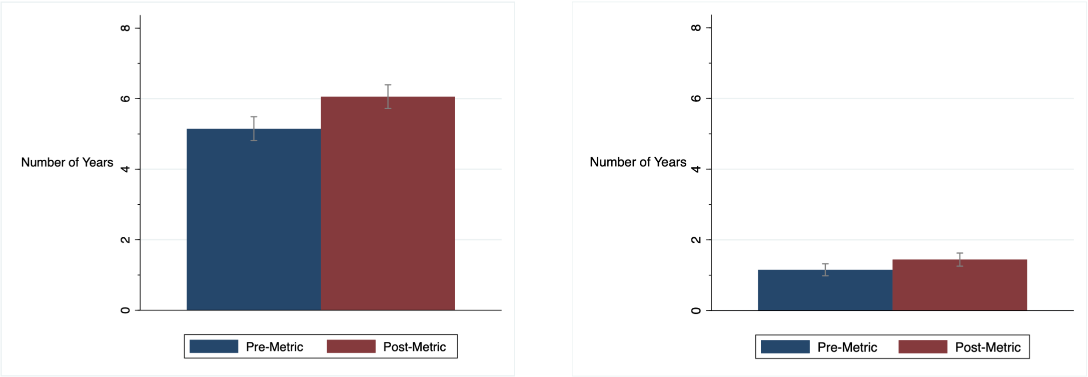
Note: The unit of analysis is the accident-occupant-vehicle level. The main sample is constructed by restricting the CDS to (1) occupants seated in the front left and second left position, (2) in passenger cars, (3) in side crash accidents. The data source is the NHTSA National Automotive Sampling System (NASS) Crashworthiness Data System (CDS) for years 1991-2007. The CDS is a representative and random sample of thousands of police-reported crashes. Field research teams based at 24 geographic sites collect data from police reports, on-sight investigations and occupant interviews. The raw number of accident-vehicle-occupants is 22,715. This corresponds to 16,573 unique occupant types (age, height, weight, gender, race) and 3,308 unique vehicle types (make, model, model year). This corresponds to 494 major generational redesigns. These vehicle types represent 46 unique makes (e.g. Toyota, Lexus, Daihatsu) owned by 24 unique parent firms (e.g. Toyota Motor Corporation).

Figure 3-1: Overview of the side impact dummy (SID)



Introduction	09/1994
Seat Position	Driver, Front Rear
Manufacturer	Humanetics
Reference Pop.	Civilian Male
Reference Size	50th Percentile
Reference Data	1960s
Weight (kg)	76.5
Height (cm)	172
Seated Height (cm)	89.9

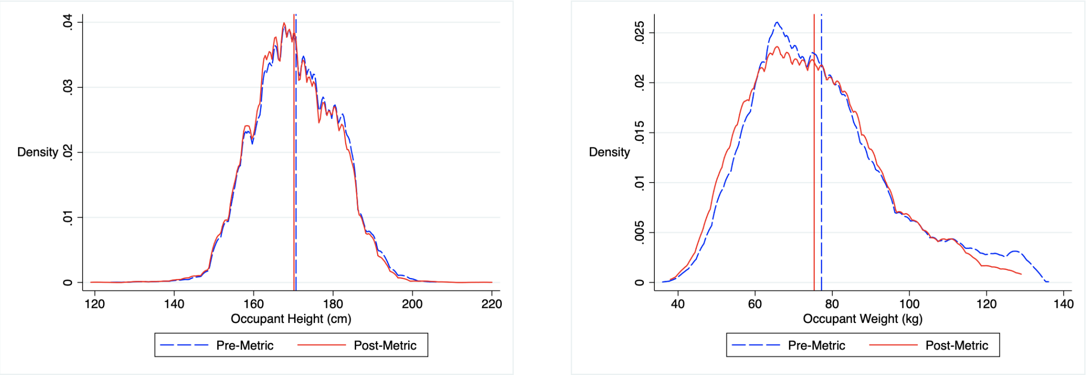
Figure 3-2: No significant change in vehicle redesign time



(a) Mean time between model-specific redesigns

(b) Mean time between firms' new/redesign models

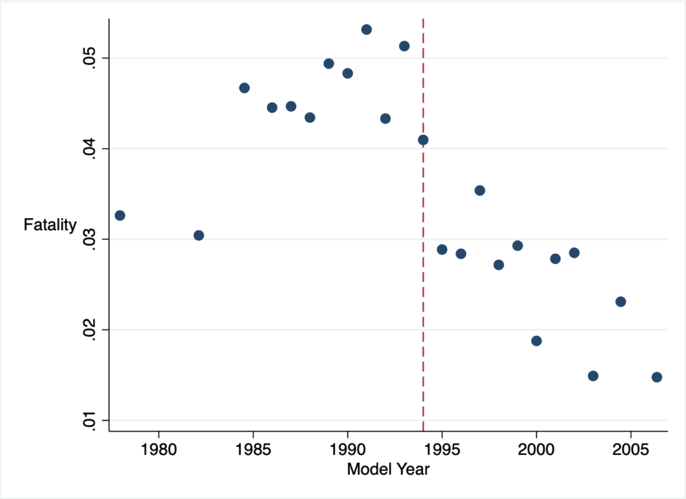
Figure 3-3: Height and weight remain stable



(a) Kernel density of occupant height

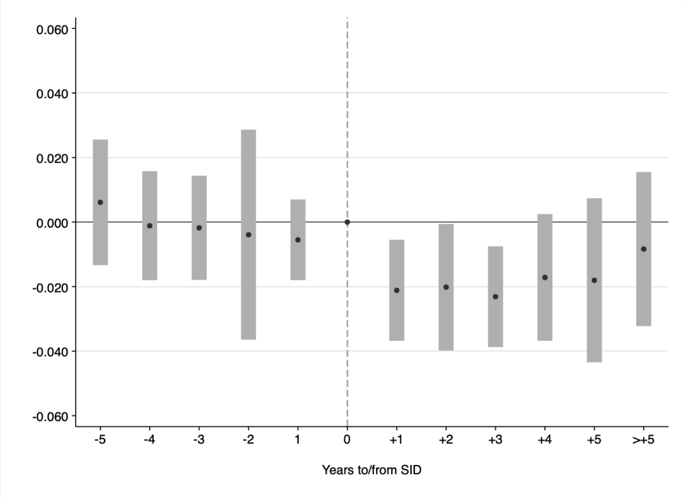
(b) Kernel density of occupant weight

Figure 3-4: Correlation between fatality rates and vehicle production year



The figure above presents a binned scatterplot that shows the correlation between the production year of the vehicle an occupant is traveling in when a crash occurs, and the probability of dying from the incident in the raw data. To construct the plot, the mean occupant fatality is plotted against 25 equal-sized bins of vehicle production year. The dashed line represents the introduction of SID in 1994.

Figure 3-5: Event study of introduction of SID on fatalities



The figure above plots estimates from a linear probability specification where fatality is regressed onto 10 interaction terms between treatment status and the number of years before/after the first post-metric major redesign. The black dots are the coefficient estimates, and the shaded grey bars are the 95% confidence interval, clustered at the make level.

Table 3.2: The introduction of the SID metric on fatalities

	1(Fatality)				
	(1)	(2)	(3)	(4)	(5)
Post-Metric Car	-0.0186***	-0.0192***	-0.0184***	-0.0183***	-0.0178***
	(0.0067)	(0.0060)	(0.0063)	(0.0063)	(0.0063)
Mean	0.036	0.036	0.036	0.036	0.036
Make× Model FE	Yes	Yes	Yes	Yes	Yes
Model Year FE	Yes	Yes	Yes	Yes	Yes
Make× Model Year FE	Yes	Yes	Yes	Yes	Yes
Accident Year FE	Yes	Yes	Yes	Yes	
Accident Controls		Yes	Yes	Yes	
Occupant Controls			Yes	Yes	
Vehicle Controls				Yes	
PDS-Lasso Controls					Yes
Observations	22,472	22,472	22,472	22,472	22,472

This table presents estimates of β_1 from equation 3.1, with the progressive addition of the controls outlined in Table 3.1. Column (1) has only make × model fixed effects, model year fixed effects, make × model year fixed effects and accident year fixed effects. Column (2) adds additional accident controls. Column (3) adds occupant controls. Column (4) adds vehicle controls. Column (5) uses the controls selected by the post-double-selection Lasso (PDS Lasso) method: occupant age squared, height, height squared, number of vehicles involved, whether the occupant was wearing a seat belt, whether there was alcohol involvement, and whether there was drug involvement, along with make-model fixed effects, make-year fixed effects, and model year fixed effects. The PDSLasso method using tools developed by Ahrens et al. (2020) based on Belloni et al. (2014). See Table 3.13 more details on the method and selected variables. Standard errors are clustered at the make level. **p<0.1, **p<0.05, ***p<0.01.*

Table 3.3: The SID metric is a focusing device

	1(Fatality)				
	(1)	(2)	(3)	(4)	(5)
Post-Metric Car	-0.0159**	-0.0166**	-0.0160**	-0.0159**	-0.0155**
	(0.0070)	(0.0062)	(0.0065)	(0.0066)	(0.0065)
Post-Metric Car	-0.0580***	-0.0577***	-0.0550***	-0.0550***	-0.0537***
× SID-like Occupant	(0.0127)	(0.0125)	(0.0121)	(0.0120)	(0.0125)
Mean	0.036	0.036	0.036	0.036	0.036
Make× Model FE	Yes	Yes	Yes	Yes	Yes
Model Year FE	Yes	Yes	Yes	Yes	Yes
Make× Model Year FE	Yes	Yes	Yes	Yes	Yes
Accident Year FE	Yes	Yes	Yes	Yes	
Accident Controls		Yes	Yes	Yes	
Occupant Controls			Yes	Yes	
Vehicle Controls				Yes	
PDS-Lasso Controls					Yes
Observations	22,472	22,472	22,472	22,472	22,472

This table presents estimates of β_1, β_3 from equation 3.2, with a progressive addition of the controls outlined in Table 3.1. SID-like Occupant is defined to be within a 0.5 standard deviation in height and weight from the Side Impact Dummy (SID) dimensions (as outlined in Figure 3-1). This is robust to more narrow and broad definitions of distance from SID dimensions as reported in Table 3.18. Column (1) has only make × model fixed effects, model year fixed effects, make × model year fixed effects and accident year fixed effects. Column (2) adds additional accident controls. Column (3) adds occupant controls. Column (4) adds vehicle controls. Column (5) uses the controls selected by the post-double-selection Lasso (PDS Lasso) method: occupant age squared, height, height squared, number of vehicles involved, whether the occupant was wearing a seat belt, whether there was alcohol involvement, and whether there was drug involvement, along with make-model fixed effects, make-year fixed effects, and model year fixed effects. The PDSLasso method using tools developed by Ahrens et al. (2020) based on Belloni et al. (2014). See Table 3.13 more details on the method and selected variables. Standard erros are clustered at the make level. Table 3.16 further splits Column 4 and 5 by gender and finds that the focus on SID-like Occupants exists in both genders. Standard errors are clustered at the make level. **p<0.1, **p<0.05, ***p<0.01.*

Table 3.4: Descriptive statistics for vehicle specifications sample

Variable	**Mean**	**Std. Dev.**	**Min.**	**Max.**
Wheelbase (in)	103.94	6.39	66.30	144.6
Length (in)	182.13	13.2	103.1	225.1
Width (in)	69.55	3.31	48.7	83.0
Height (in)	55.34	3.86	43.5	90.0
Curb Weight (lbs)	3097.88	534.28	1620.0	5413.0
Net Horsepower	175.296	68.45	49	605
Miles Per Gallon	29.14	5.43	12	68
Model Year	1998	4.85	1991	2006
Price	25551.75	17602.10	4825	443000
Major Redesign	0.3	0.458	0	1

Note: The unit of analysis is the vehicle make-model-model year level. The data source is the WARD's Automotive Yearbook for year 1991-2006. The sample restricts the Yearbook data to passenger cars without optional upgrades. Major Redesign is equal to 1 if the model has just undergone a major redesign or is being launched in that year. Data on major redesigns were hand-collected from the industry summary sections of WARD's Automotive Yearbooks, and supplemented with web searches on Wikipedia and KellyBlueBook.

Table 3.5: Improving on SID required meaningful changes

	(1) Width	(2) Weight	(3) Height	(4) Wheelbase	(5) Length	(6) Horsepower	(7) Miles/Gallon
Post × Redesign	0.169**	38.956***	0.062	-0.024	0.771**	0.201	-0.305**
	(0.084)	(9.283)	(0.081)	(0.132)	(0.320)	(0.981)	(0.151)
Make-Model FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Car Type FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	6,573	6,552	6,573	6,573	6,575	6,572	6,524

This table presents estimates of the impact of undergoing a major redesign after the introduction of SID on vehicle design choices. Vehicle make-model fixed effects and car type (e.g. 4-door sedan, 3-door hatchback, etc.) are included. Robust standard errors are reported. *p<0.1, **p<0.05, ***p<0.01

Table 3.6: Increasing vehicular width and weight is correlated with lower fatalities

	1(Fatality)	
	(1) Accid-Occ-Veh	(2) PDS-Lasso
Width (in)	-0.0015*** (0.0004)	-0.0014*** (0.0005)
Weight (10 lbs)	-0.0010*** (0.0002)	-0.0011*** (0.0003)
Length (in)	-0.0004*** (0.0001)	-0.0004*** (0.0001)
Miles/Gallon	0.0009** (0.0003)	0.0009*** (0.0003)
Make× Model FE	Yes	Yes
Model Year FE	Yes	Yes
Make× Model Year FE	Yes	Yes
Observations	22,530	22,530

This table presents estimates of the relationship between vehicle design characteristics, and the likelihood of incurring a fatality in a side impact crash. Column 1 presents estimates using the accident, occupant, and vehicle controls listed in Table 3.1. Column 2 uses the controls selected by the post-double-selection lasso method detailed in Table 3.13. Standard errors are clustered at the vehicle make level. **p<0.1, **p<0.05, ***p<0.01*

Table 3.7: Firms with safety positioning have broader focus

	1(Fatality)			
	(1)	(2)	(3)	(4)
	Annual Reports		Media News	
	Low	High	Low	High
Post-Metric Car	-0.0089	-0.0240***	-0.0084	-0.0238***
	(0.0074)	(0.0031)	(0.0072)	(0.0057)
Post-Metric Car	-0.0530***	-0.0579**	-0.0524***	-0.0585***
× SID-like Occupant	(0.0115)	(0.0215)	(0.0147)	(0.0187)
Mean	0.036	0.037	0.034	0.037
Make × Model FE	Yes	Yes	Yes	Yes
Model Year FE	Yes	Yes	Yes	Yes
Accident Year FE	Yes	Yes	Yes	Yes
Observations	9,023	12,278	9,546	13,005

This table presents estimates of $\beta_1 \beta_3$ from the equation 3.2 split by subsamples. Columns 1 and 2 split the sample by firms with below and above the median fraction of safety-related language in their annual shareholders reports in the years preceding SID respectively. Columns 3 and 4 split the sample by firms with below and above the median fraction of safety-related language in the media coverage of the new product launches. All regression include the controls outlined in Table 3.1. SID-like Occupant is defined to be within a 0.5 standard deviation in height and weight from the Side Impact Dummy (SID) dimensions (as outlined in Figure 3-1). Standard errors are clustered at the parent level for shareholder reports, and make level for media news coverage. **p<0.1, **p<0.05, ***p<0.01.*

Table 3.8: Firms with prior knowledge have broader focus

	1(Fatality)			
	(1)	(2)	(3)	(4)
	Safety Patents		Safety Defects	
	Low	High	More	Less
Post-Metric Car	-0.0099	-0.0208***	-0.0113	-0.0227***
	(0.0070)	(0.0011)	(0.0067)	(0.0049)
Post-Metric Car	-0.0642***	-0.0414**	-0.0680***	-0.0498**
× SID-like Occupant	(0.0164)	(0.0087)	(0.0116)	(0.0179)
Mean	0.037	0.034	0.037	0.036
Make × Model FE	Yes	Yes	Yes	Yes
Model Year FE	Yes	Yes	Yes	Yes
Accident Year FE	Yes	Yes	Yes	Yes
Observations	10,252	10,911	10,432	12,233

This table presents estimates of β_1, β_3 from the equation 3.2 split by subsamples. Columns 1 and 2 split the sample by firms with below and above the median number of patents that have safety-related language respectively. Columns 3 and 4 split the sample by firms with above and below the median proportion of defects investigations opened relative to their total vehicle models on the road respectively. All regression include the controls outlined in Table 3.1. SID-like Occupant is defined to be within a 0.5 standard deviation in height and weight from the Side Impact Dummy (SID) dimensions (as outlined in Figure 3-1). Standard errors are clustered at the parent level for patents, and make level for defects investigations. **p<0.1, **p<0.05, ***p<0.01.*

Table 3.9: Prior knowledge mitigates metric myopia

	1(Fatality)			
	(1)	(2)	(3)	(4)
	Low Safety Positioning		High Safety Positioning	
	Low Knowledge	High Knowledge	Low Knowledge	High Knowledge
Post-Metric Car	-0.0051	-0.0205	-0.0263*	-0.0201**
	(0.0071)	(0.0119)	(0.0114)	(0.0005)
Post-Metric Car	-0.0514**	-0.0455	-0.0912**	-0.0358
× SID-like Occupant	(0.0159)	(0.0106)	(0.0259)	(0.0093)
Make × Model FE	Yes	Yes	Yes	Yes
Model Year FE	Yes	Yes	Yes	Yes
Accident Year FE	Yes	Yes	Yes	Yes
Observations	6,742	2,330	3,508	8,580

This table presents estimates of β_1, β_3 from the equation 3.2 split by subsamples. Columns 1 and 2 examine firms with below the median safety-related language in their annual reports, further split into those with low and high safety patents respectively. Columns 3 and 4 examine firms with above the median safety-related language, and similarly splits it into low and high safety patents. All regression include the controls outlined in Table 3.1. SID-like Occupant is defined to be within a 0.5 standard deviation in height and weight from the Side Impact Dummy (SID) dimensions. Standard errors are clustered at the parent level. **p<0.1, **p<0.05, ***p<0.01.*

Table 3.10: Prior knowledge overcomes the trade-off

	Miles/Gallon			
	(1)	(2)	(3)	(4)
	Safety Patents		Safety Defects	
	Low	High	More	Less
Post × Redesign	-0.955***	-0.017	-0.871***	0.300
	(0.207)	(0.197)	(0.237)	(0.197)
Make-Model FE	Yes	Yes	Yes	Yes
Car Type FE	Yes	Yes	Yes	Yes
Observations	2,868	3,998	3,836	3,031

This table presents estimates of the impact of undergoing a major redesign after the introduction of SID on miles per gallon. Vehicle make-model fixed effects and car type (e.g. 4-door sedan, 3-door hatchback, etc.) are included. Columns 1 and 2 split the sample by firms with below and above the median number of patents that have safety-related language respectively. Columns 3 and 4 split the sample by firms with above and below the median proportion of defects investigations opened relative to their total vehicle models on the road respectively. Robust standard errors are reported. **p<0.1, **p<0.05, ***p<0.01*

Sorted Partial Effects Analysis

The main analysis (presented in Table 3.2) reports the average partial effect (i.e. treatment effect) from a linear probability model of the introduction of SID to a vehicle model on the likelihood of occupant fatality, holding production year, vehicle model, occupant, and accident characteristics constant. While this is the best linear predictor of the treatment effect, by collapsing the full range of marginal partial effects into one summary statistic, a lot of potentially informative heterogeneity is lost.

To explore this heterogeneity, I relax the assumption of linearity to explore how SID, as a new metric for side impact safety, may have impacted some occupants differently from others as a complement to my main analysis. To do so I estimate a model that fully interacts a dummy for the introduction of SID to a vehicle model —the treatment —with all the occupant, vehicle and accident characteristics as outlined in Table 3.1:

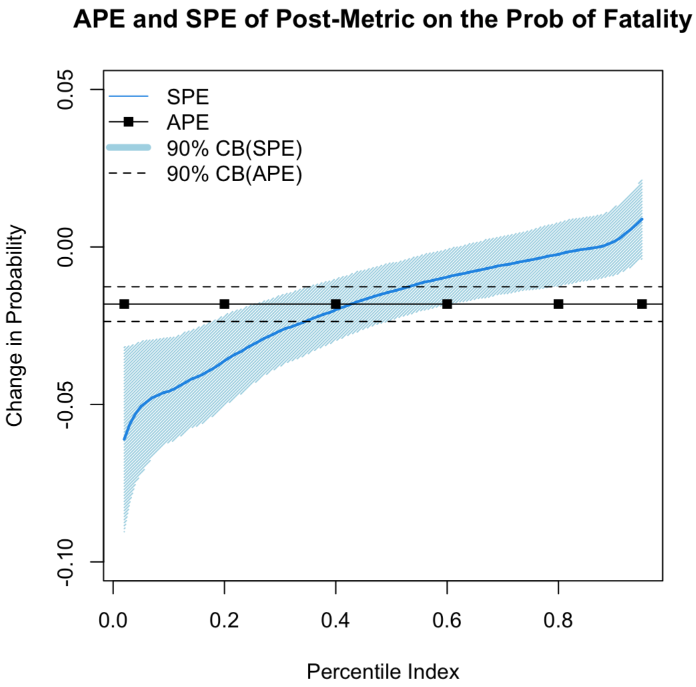
$$E[\text{Fatal}_i | X_{ik}, \text{PostMetric}_{jt}] = \beta_0 + \beta_1 \text{PostMetric}_{jt} \times (X_{ik}, \delta_j, \gamma_t) + \epsilon_{ijtk} \tag{3.4}$$

I employ the sorted partial effects method (SPE) for interactive linear models to summarize the results following Chernozhukov et al. (2018). SPE involves computing the partial effects of SID on the likelihood of fatality at different quantiles of the partial effect, sorting them in order from lowest to highest, and plotting them. This supplements the main analysis by visualizing the full range of heterogeneity in partial effects, and identifying which occupant, vehicle and accident characteristics are associated with the most strong or least strong effects.

Figure 3-6 plots the estimated partial effects from the fully interacted model in Equation 3.4, sorted and indexed by quantiles. The sorted effects are plotted in blue while the average partial effect is in black. In this case, the conditional quantile is estimated for the 0.02 to 0.98 quantiles. The quantiles are plotted along the X-axis and the predicted impact on fatalities is on the Y-axis, such that for example, at x = 0.2 there is a 20% chance the actual decline in fatalities is larger than the prediction, and an 80% chance it is smaller. This shows the probability of fatality ranges from -0.072 to +0.0043. In other words, there exists a subgroup of occupants for whom the introduction of the metric decreased their likelihood of fatality approximately twice as much as the average, and a subgroup for whom they are actually worse off after the introduction of SID.

The table in Figure 3-6 summarizes the corresponding classification analysis which presents the mean and standard error for the characteristics of control variables in the 10% most and least affected groups in terms of post-SID fatality rate reductions. This finds that relative to accident characteristics, occupant characteristics drive this gap. The group with the largest gains from the introduction of the metric, have a median height and weight that tracks closely to that of SID. I also find that occupants in this group tend to me older and male which corresponds to the fact that SID was developed based on data from earlier cadaver crash tests.

Figure 3-6: Sorted effects of Post-SID cars on probability of fatality



Variable	10% Least Improved	SE	10% Most Improved	SE
Age	21.17	1.49	59.79	5.12
Height	167.70	2.12	172.73	1.50
Weight	71.52	5.66	77.37	2.60
Male	0.29	0.11	0.66	0.08
Alcohol	0.01	0.01	0.05	0.17
Drug	0.00	0.05	0.09	0.09
No Seatbelt	0.29	0.15	0.14	0.07

This above figure depicts the estimated partial effects, sorted and indexed by percentiles, from a regression model where an indicator variable for Post-Metric vehicle model is fully interacted with all accident, occupant and vehicle controls as outlined in Table 3.1. The blue line plots the level of heterogeneity in the improvements from Post-Metric car, and the dotted black line plots the average partial effects. 95% bootstrap uniform confidence bands are shaded in blue. Below, the table summarizes the corresponding classification analysis which summarizes the characteristics of the most versus least affected groups. The analysis and figure were implemented using the spe package in R developed by Chen et al. (2019) based on Chernozhukov et al. (2018).

Supervised ML for Text Classification

In order to understand the degree to which firms had developed safety-oriented positioning and capabilities, several different data sources were used as paper trails for these strategic investments: shareholders reports, media coverage, and patent text. Given that these are sources of text data it was necessary to classify text as being safety-oriented versus not.

I employed a standard approach to text pre-processing by normalizing it to lowercase, lemmatizing, and remove special characters, HTML tags, and extra white space. Then, I used two different approaches to classify the text, which generated similar classification outcomes.

The first was a straightforward frequency count of the number of safety-related keywords in the text. Safety keywords were obtained from the NHTSA Glossary Of Highway Safety Terms and Definitions, excluding those related to NHTSA-specific acronyms or programs (e.g. FAIN, final financial reconciliation; Hatch Act, etc.) or broad non-safety terms (e.g. internal controls, countermeasure).[27] This was also compared with the keywords extracted by Giorgi & Edward (2019), which employed topic modeling analysis on automotive safety legislation texts from 1965-1980 (including congressional hearings, advertising copy, recordings of meetings between President Nixon and automotive firm executives, NHTSA analyses and letters, congressional reports) to ensure sufficient coverage of safety-related keywords. This was then weighted against the total number of words in the text to generate the fraction of words in the text that are safety keywords, allowing for the creation of an indicator variable for being above the median in safety frequency.

[27] See https://www.nhtsa.gov/resources-guide/glossary-highway-safety-terms-and-definitions for the full glossary.

The second approach was to train a supervised machine learning model using the scikit-learn and nltk libraries in Python. To train and test the model, data from the Society of Automotive Engineers (SAE) ground vehicle articles were used. The SAE is a trade association that publishes technical articles, best practices and recommended voluntary standards covering mobility engineering in aerospace, commercial and motor vehicles. To date, it has published over 10,000 documents regarding proposed standards that are written and maintained by 9,000 volunteer engineers.[28]. To create the sample, all ground vehicle proposed standards and technical reports published in the years 1970 to 1993 from SAE Mobilus were collected. Those tagged as safety related, issued by the Vehicle Safety Systems subcommittees, or reference by NHTSA policies[29] were classified as being safety-related.

Preprocessed text was converted into numerical features using the bag of words model. The values outputted from this model were then converted into term frequency-inverse document frequency (TF-IDF) values. The sample was then randomly split into a training set (67%) and a test set (33%) to build and assess the accuracy of the binary text classification model in determining if text was safety-related or not. A naïve Bayes classifier, decision tree classifier, and maximum entropy classifier were evaluated against one another. The naïve Bayes generated the highest accuracy at 0.80 and was the classification approach used.

[28] https://www-sae-org.libproxy.mit.edu/servlets/works/customer.do

[29] A list of documents that were referenced by different US government departments was provided to the author by the SAE.

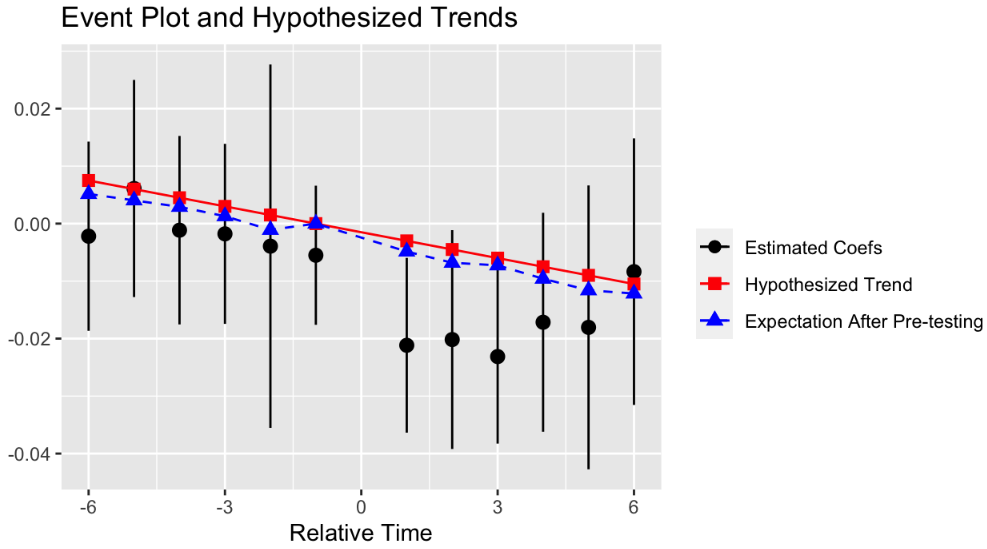
Additional Difference-in-Differences Analyses

This section employs recent advances in econometrics to conduct additional analyses of the difference-in-differences (DiD) approach used in the main analysis of the paper.

The lack of pre-trends —"parallel trends" —is a core assumption of the DiD approach, and in the main analysis, I employ the standard approach of using an event study plot to visually inspect for pre-trends in Figure 3-5. I find that for the periods leading up to the treatment, the introduction of the SID metric, the estimated coefficients are statistically insignificant from zero. Recent research however, finds that this approach may be low powered and unable to detect violations of the parallel trends assumption (Freyaldenhoven et al., 2019; Roth, 2020). For instance, a key concern could be that there are pre-trends that fit within the confidence interval (and are similarly powered to evidence of no pre-trends).

To mitigate this concern, I employ a diagnostic approach following Roth (2020) in which I visualized different hypothesized linear violations of pre-trends, and computed the chance that it would be detected as significant. The general approach was to hypothesize the case where fatality rates were on the decline even absent the introduction of SID, and modifying the parameters (e.g. slope) to get closer to the estimates in the event study plot until power had been minimized. Figure 3-7 presents the hypothesized trend with the case with the lowest power found in my analysis (in red). The blue dashed line plots the expected estimates conditional on passing the pre-trends test under the hypothesized trend. In this extreme case, the probability that the event study analysis would detect a significant pre-trend is 0.325.

Figure 3-7: Analysis of event study versus potential, hypothesized pre-trends



Appendix Tables and Figures

Table 3.11: Summary of main data sources

Data Source	Provider	Data Type
Crashworthiness Data System	NHTSA	Representative, random sample of US passenger vehicle crashes
Ward's Automotive Yearbook	WardsAuto	Industry publication covering specifications for US-sold vehicles
Mergent Archives	FTSE Russell	Historical data on public companies including annual report filings
PatentsView	USPTO	US patent applications and grants, inventors, organizations and locations
Factiva	Dow Jones	US media news articles covering new products and product launches

Table 3.12: Characteristics of the main crash sample

Variable	Raw Count
Number of accidents-vehicle-occupants	22,715
Number of accidents	17,216
Number of unique occupant types *(age, height, weight, male, race)*	16,573
Number of unique vehicle types *(make, model, model year)*	3,308
Number of unique makes (e.g. Toyota, Lexus, Daihatsu)	46
Number of unique parent firms (e.g. Toyota Motor Corporation)	24

Note: The data source is the NHTSA National Automotive Sampling System (NASS) Crashworthiness Data System (CDS) for years 1991-2007. The CDS is a representative and random sample drawn from all crashes in the United States that (1) are police-reported, (2) involve a harmful event of property damage or personal injury, (3) involve a towed passenger car, light truck, or van on a traffic way. Field research teams with trained crash investigators at 24 geographic sites collect data from police reports, and conduct on-sight investigations and occupant interviews. For the main analysis, the sample restricts the full CDS data to (1) occupants seated in the front left and second left position, (2) in passenger cars, (3) in side crash accidents.

Table 3.13: Post double selection lasso variables

Double-selected Variables	Post-Lasso OLS Coefficient
Age^2	0.00001
	(0.00000)
Height^2	-0.000001
	(0.00000)
Height	0.00036
	(0.00019)
Number of Vehicles	-0.01760
	(.00276)
Alcohol Involved	0.04707
	(0.00796)
Drugs Involved	0.07950
	(0.01357)
No Seat Belt	0.03516
	(0.00589)

To supplement the analysis using accident, vehicle and occupant controls selected by the author, the post double selection lasso (PDS-Lasso) method developed by Belloni et al. (2014) was used. PDS-Lasso uses lasso to optimally select regression control variables to mitigate over-fitting and omitted variable bias. This involves three steps: (1) a lasso regression with the dependent variable, Fatality, as the outcome of interest, and all potential control variables; (2) a lasso regression with the focal independent variable, Post-SID Car, as the outcome of interest and all potential controls; and (3) a linear regression with Fatality on Post-SID Car including the control variables in the lasso regressions that had non-zero coefficients. The potential set of regressors included 102 variables, including all of those listed in Table 3.1 and the squares of those variables, as well as additional controls such as the number of vehicles, number of occupants, pregnancy, day of the week, speed limit, drive race dummies, weather condition dummies, impact angle, vehicle body type, other vehicle's weight, number of lanes, etc.

Table 3.14: Estimates for Equation 3.1 using probit regression

	1(Fatality)			
	(1)	(2)	(3)	(4)
Post-Metric Car	-0.2927***	-0.3308***	-0.3193***	-0.3194***
	(0.0807)	(0.0760)	(0.0847)	(0.0849)
Make × Model FE	Yes	Yes	Yes	Yes
Model Year FE	Yes	Yes	Yes	Yes
Accident Year FE	Yes	Yes	Yes	Yes
Accident Controls		Yes	Yes	Yes
Occupant Controls			Yes	Yes
Vehicle Controls				Yes
Observations	20,523	20,523	20,523	20,523

This table presents estimates of β_1 from equation 3.1 using a probit regression, with the progressive addition of the controls outlined in Table 3.1. Column 1 has make-model fixed effects, model year fixed effects and accident year fixed effects. Column 2 adds accident controls. Column 3 adds occupant controls. Column 4 adds occupant controls. Standard errors clustered at the make level are reported. **p<0.1, **p<0.05, ***p<0.01.*

Table 3.15: Estimates for Equation 3.2 using probit regression

	1(Fatality)			
	(1)	(2)	(3)	(4)
Post-Metric Car	-0.2672***	-0.3062***	-0.2990***	-0.2991***
	(0.0821)	(0.0780)	(0.0851)	(0.0853)
Post-Metric Car	-0.5499***	-0.5207***	-0.4323**	-0.4327**
× SID-like Occupant	(0.1783)	(0.1857)	(0.1926)	(0.1927)
Make × Model FE	Yes	Yes	Yes	Yes
Model Year FE	Yes	Yes	Yes	Yes
Accident Year FE	Yes	Yes	Yes	Yes
Accident Controls		Yes	Yes	Yes
Occupant Controls			Yes	Yes
Vehicle Controls				Yes
Observations	20,523	20,523	20,523	20,523

This table presents estimates of β_1, β_3 from equation 3.2 using a probit regression, with the progressive addition of the controls outlined in Table 3.1. Column 1 has make-model fixed effects, model year fixed effects and accident year fixed effects. Column 2 adds accident controls. Column 3 adds occupant controls. Column 4 adds occupant controls. Standard errors clustered at the make level are reported. **p<0.1, **p<0.05, ***p<0.01.*

Table 3.16: SID-like Occupants results are present for both genders

	1(Fatality)			
	(1)	(2)	(3)	(4)
Controls	Accid-Occ-Veh		PDSLasso	
Sample	Male	Female	Male	Female
Post-Metric Car	-0.0212*	-0.0142	-0.0225*	-0.0138
	(0.0112)	(0.0141)	(0.0120)	(0.0139)
Post-Metric Car× SID-like Occupant	-0.0274*	-0.1192***	-0.0252*	-0.1162***
	(0.0145)	(0.0282)	(0.0144)	(0.0283)
Mean	0.040	0.0327	0.040	0.0327
Make× Model FE	Yes	Yes	Yes	Yes
Model Year FE	Yes	Yes	Yes	Yes
Make× Model Year FE	Yes	Yes	Yes	Yes
Accident Year FE	X	X		
Observations	11,228	11,109	11,228	11,109

This table presents estimates of β_1, β_3 in equation 3.2 in subsamples split by gender. Columns (1) and (2) include accident, occupant, and vehicle controls. Columns (3) and (4) includes the PDSLasso controls described in Table 3.13. Vehicle make × model, vehicle make× model year and model year fixed effects are included in all regressions. Standard errors are clustered at the vehicle make level. **p<0.1, **p<0.05, ***p<0.01*

Table 3.17: Estimates for Equations 3.1 and 3.2 using injury severity

	1(Fatality)			
	(1)	(2)	(3)	(4)
Post-Metric Car	-0.0857***	-0.0781***	-0.0797***	-0.0732***
	(0.0294)	(0.0287)	(0.0289)	(0.0284)
Post-Metric Car			-0.1629**	-0.1360*
× SID-like Occupant			(0.0714)	(0.0714)
Make × Model FE	Yes	Yes	Yes	Yes
Model Year FE	Yes	Yes	Yes	Yes
Accid-Occ-Veh Controls	Yes	No	Yes	No
PDS-Lasso Controls	No	Yes	No	Yes
Observations	22,435	22,435	22,435	22,435

This table presents estimates of Equations 3.1 and 3.2 using ordered probit regressions. Columns 1 and 3 include the controls outlined in Table 3.1. Columns 2 and 4 use post-double selection LASSO controls. Standard errors clustered at the make level are reported. Injury severity is an ordinal variable ranging from 0 to 4 based on police reports. *p<0.1, **p<0.05, ***p<0.01.

Table 3.18: Equation 3.2 using alternative definitions of SID-like Occupant

	1(Fatality)				
	(1)	(2)	(3)	(4)	(5)
SID-like Occupant	0.7xSD	0.6xSD	0.5xSD	0.4xSD	0.3xSD
Post-Metric Car	-0.0160**	-0.0156**	-0.0158**	-0.0158**	-0.0160**
	(0.0067)	(0.0066)	(0.0066)	(0.0066)	(0.0066)
Post-Metric Car ×SID-like Occupant	-0.0312***	-0.0516***	-0.0547***	-0.0547***	-0.0507***
	(0.0110)	(0.0099)	(0.0122)	(0.0122)	(0.0127)
Mean	0.036	0.036	0.036	0.036	0.036
Make× Model FE	Yes	Yes	Yes	Yes	Yes
Model Year FE	Yes	Yes	Yes	Yes	Yes
Make× Model Year FE	Yes	Yes	Yes	Yes	Yes
Accident Year FE	Yes	Yes	Yes	Yes	Yes
Observations	22,551	22,551	22,551	22,551	22,551

This table presents estimates of β_1, β_3 in equation 3.2 using alternative definitions of an occupant being SID-like. Columns (1) through (5) define SID-like Occupant as $X \times \sigma$ where X varies from 0.3 to 0.7. The standard deviation, σ, of occupant height is 10.168 cm and occupant weight is 17.855 kg (as summarized in Table 3.1).

Table 3.19: CEM estimates for Equations 3.1 and 3.2

	1(Fatality)	
	(1)	(2)
Post-Metric Car	-0.0188***	-0.0175**
	(0.0054)	(0.0054)
Post-Metric Car		-0.0419**
× SID-like Occupant		(0.0156)
Observations	21,604	21,604

This table presents estimates of Equations 3.1 and 3.2 using coarsened exact matching. **p<0.1, **p<0.05, ***p<0.01.*

Table 3.20: Results robust to narrowing sample period by vehicle age

	1(Fatality)					
	(1)	(2)	(3)	(4)	(5)	(6)
	≤ 15 years		≤ 10 years		≤ 5 years	
Post-Metric Car	-0.0198***	-0.0176***	-0.0195***	-0.0177***	-0.0120*	-0.0108
	(0.0044)	(0.0045)	(0.0047)	(0.0046)	(0.0071)	(0.0072)
Post-Metric Car		-0.0600***		-0.0487***		-0.0349*
× SID-like Occupant		(0.0117)		(0.0163)		(0.0198)
Make× Model FE	Yes	Yes	Yes	Yes	Yess	Yes
Model Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	21,471	21,471	17,504	17,504	11,028	11,028

This table presents estimates of Equations 3.1 and 3.2 with the sample being progressively restricted further by vehicle age (computed as Accident Year - Model Year). All regressions include the controls outlined in Table 3.1. Standard errors are clustered at the make level. $^{*}p<0.1$, $^{**}p<0.05$, $^{***}p<0.01$.

Table 3.21: Results robust to narrowing sample by model years

	1(Fatality)			
	(1)	(2)	(3)	(4)
	1985-2005		1990-2000	
Post-Metric Car	-0.0188***	-0.0171***	-0.0198***	-0.0187***
	(0.0048)	(0.0048)	(0.0055)	(0.0055)
Post-Metric Car		-0.0475***		-0.0358**
× SID-like Occupant		(0.0125)		(0.0161)
Make × Model FE	Yes	Yes	Yes	Yes
Model Year FE	Yes	Yes	Yes	Yes
Observations	19,616	19,616	11,341	11,341

This table presents estimates of Equations 3.1 and 3.2 with the sample being progressively restricted further by the model years of the vehicles. All regressions include the controls outlined in Table 3.1. Standard errors are clustered at the make level. **p<0.1, **p<0.05, ***p<0.01.*

Table 3.22: Results robust to narrowing sample to major redesigns

	1(Fatality)			
	(1)	(2)	(3)	(4)
Post-Metric Car	-0.0249***	-0.0236***	-0.0256***	-0.0244***
	(0.0066)	(0.0067)	(0.0078)	(0.0079)
Post-Metric Car		-0.0431*		-0.0404*
× SID-like Occupant		(0.0241)		(0.0237)
Make × Model FE	Yes	Yes	Yes	Yes
Observations	5,062	5,062	5,062	5,062

This table presents estimates of Equations 3.1 and 3.2 narrowing the subsample to only vehicles in the produced in the year they are redesigned. Columns 1 and 3 include the controls outlined in Table 3.1. Columns 2 and 4 use post-double selection LASSO controls. Standard errors clustered at the make level are reported. Injury severity is an ordinal variable ranging from 0 to 4 based on police reports. **p<0.1, **p<0.05, ***p<0.01.*

Table 3.23: Summary of main and alternative samples

Sample	Sampling Method	Type of Accidents	Occ Height/Weight Avail	Primary Data Source
NASS-CDS	Random within representative stratum	Police-reported, Harmful event, Passenger/Van/Light Vehicle	All Occupants	On-site investigations, interviews, police reports
FARS	Universe	1+ Fatality Occurs	Driver Only (1998-)	Police reports, coroner/death certificates, state DOT data
FARS-LP		*Fatality in other vehicle*		
GES	Random within representative stratum	Police-reported, Harmful event	None	Police reports
GES-BRFSS			*Imputed for drivers*	

NASS-CDS, FARS and GES are abbreviations for the National Automotive Sampling System - Crashworthiness Data System, Fatality Analysis Reporting System, and General Estimates Survey System respectively. All three data sources are compiled and distributed by the National Highway Traffic Safety Administration, part of the US Department of Transportation. NASS-CDS is the main sample used in this study. FARS-LP refers to the sample restriction approach developed by Levitt and Porter (2001) to mitigate selection issue of fatality-only accidents bering recorded in the FARS sample: focusing only on 2-passenger car accidents in which a fatality occurs in the other vehicle. GES+BRFSS refers to the use of the Behavioral Risk Factor Surveillance System (BRFSS) compiled and distributed by the Centres for Disease Control and Prevention. The BRFSS was used to impute the height and weight of the driver based on the accident year and the age, region, race and sex of the driver using the mi package in Stata.

Table 3.24: Descriptive statistics for the FARS and GES sample

Sample	FARS (N= 28065)				GES (N = 393918)			
Variable	**Mean**	**Std. Dev.**	**Min.**	**Max.**	**Mean**	**Std. Dev.**	**Min.**	**Max.**
Accident Controls								
Accident Year	2003	3.053	1998	2008	2000	4.621	1990	2006
Accident Month	6.711	3.401	1	12	6.573	3.469	1	12
Speed Limit	49.339	11.252	0	75	40.674	11.219	0	75
Wet Conditions	0.174	0.379	0	1	0.204	0.403	0	1
Dark Conditions	0.308	0.461	0	1	0.142	0.349	0	1
Daytime	0.692	0.461	0	1	0.858	0.349	0	1
Alcohol Involved	0.079	0.269	0	1	0.013	0.114	0	1
Drugs Involved	0.014	0.118	0	1	0.002	0.043	0	1
Occupant Controls								
Age	37.845	16.141	2	96	38.033	16.773	1	102
Male	0.704	0.457	0	1	0.601	0.49	0	1
Height (cm)	174.033	10.006	91.44	213.36	172.279	10.135	140.619	213.973
Weight (kg)	78.932	18.402	25.397	225.85	78.407	18.32	22.693	168.686
No Seat Belt	0.156	0.363	0	1	0.067	0.25	0	1
Vehicle Controls								
Model year	1995	6.293	1966	2009	1993	6.521	1974	2007

Table 3.25: Estimates for Equation 3.1 using alternative samples

	1(Fatality)		
	(1) FARS-LP	(2) FARS-S	(3) GES-BRFSS
Post-Metric Car	-0.0131*** (0.0024)	-0.0140** (0.0064)	-0.0015*** (0.0002)
Mean	0.0362	0.0656	0.0042
Make × Model FE	Yes	Yes	Yes
Accident Year FE	Yes	Yes	Yes
Years Covered	1998-2007	1998-2007	1991-2007
Observations	28,055	7,754	39,3918

The table presents estimates of β_1 in equation 3.1 both in the main sample and alternative data sources. Column 1 uses the FARS sample following the Levitt and Porter (2001) approach. Column 2 further restricts the FARS sample to vehicles that are struck in a side crash. Column 3 uses the GES sample with occupant heights and weights generated using multiple imputation using BRFSS data. Standard errors are clustered at the make level.

Table 3.26: Estimates for Equation 3.2 using alternative samples

	1(Fatality)		
	(1) FARS-LP	(2) FARS-S	(3) GES-BRFSS
Post-Metric Car	-0.0127***	-0.0127**	-0.0010***
	(0.0023)	(0.0064)	(0.0002)
Post-Metric Car	-0.0313*	-0.1123*	-0.0048***
× SID-like Occupant	(0.0172)	(0.0666)	(0.0018)
Mean	0.0362	0.0656	0.0042
Make × Model FE	Yes	Yes	Yes
Accident Year FE	Yes	Yes	Yes
Years Covered	1998-2007	1998-2007	1991-2007
Observations	28,055	7,754	39,3918

The table presents estimates of β_1, β_3 from equation 3.2 both in the main sample and alternative data sources. Column 1 uses the FARS sample following the Levitt and Porter (2001) approach. Column 2 further restricts the FARS sample to vehicles that are struck in a side crash. Column 3 uses the GES sample with occupant heights and weights generated using multiple imputation using BRFSS data. Standard errors are clustered at the make level.

Table 3.27: Falsification test with frontal and rear accidents

	1(Fatality)			
	(1)	(2)	(3)	(4)
Post-Metric Car	-0.0000	-0.0000	-0.0020	-0.0020
	(0.0120)	(0.0118)	(0.0121)	(0.0119)
Post-Metric Car		0.0004		-0.0011
× SID-like Occupant		(0.0233)		(0.0231)
Make × Model FE	Yes	Yes	Yes	Yes
Make × Model Year FE	Yes	Yes	Yes	Yes
Model Year FE	Yes	Yes	Yes	Yes
Accid-Occ-Veh Controls	Yes	No	Yes	No
PDS-Lasso Controls	No	Yes	No	Yes
Observations	20,514	20,514	20,514	20,514

This table presents a falsification test that estimates Equations 3.1 and 3.2 on a sample of non-side (frontal and rear-end) accidents. Columns 1 and 3 include the controls outlined in Table 3.1. Columns 2 and 4 use post-double selection LASSO controls. Standard errors clustered at the make level are reported. **p<0.1, **p<0.05, ***p<0.01.*

Table 3.28: Falsification test with seats unaffected by SID

	1(Fatality)			
	(1)	(2)	(3)	(4)
Post-Metric Car	0.0004	0.0001	0.0002	-0.0001
	(0.0153)	(0.0154)	(0.0162)	(0.0163)
Post-Metric Car		0.0184		0.0163
× SID-like Occupant		(0.0316)		(0.0317)
Make × Model FE	Yes	Yes	Yes	Yes
Make × Model Year FE	Yes	Yes	Yes	Yes
Model Year FE	Yes	Yes	Yes	Yes
Accid-Occ-Veh Controls	Yes	No	Yes	No
PDS-Lasso Controls	No	Yes	No	Yes
Observations	20,514	20,514	20,514	20,514

This table presents a falsification test that estimates Equations 3.1 and 3.2 on a sample of occupants in seats unaffected by SID. Columns 1 and 3 include the controls outlined in Table 3.1. Columns 2 and 4 use post-double selection LASSO controls. Standard errors clustered at the make level are reported. **p<0.1, **p<0.05, ***p<0.01.*

Table 3.29: Falsification test using non-fatal injury as the outcome

	1(Fatality)			
	(1)	(2)	(3)	(4)
Post-Metric Car	-0.0199	-0.0203	-0.0210	-0.0218
	(0.0135)	(0.0135)	(0.0138)	(0.0138)
Post-Metric Car			0.0253	0.0294
× SID-like Occupant			(0.0268)	(0.0285)
Make × Model FE	Yes	Yes	Yes	Yes
Make × Model Year FE	Yes	Yes	Yes	Yes
Model Year FE	Yes	Yes	Yes	Yes
Accid-Occ-Veh Controls	Yes	No	Yes	No
PDS-Lasso Controls	No	Yes	No	Yes
Observations	22,551	22,551	22,551	22,551

This table presents a falsification test that estimates Equations 3.1 and 3.2 using non-fatal major injury as the dependent variable. Columns 1 and 3 include the controls outlined in Table 3.1. Columns 2 and 4 use post-double selection LASSO controls. Standard errors clustered at the make level are reported. **p<0.1, **p<0.05, ***p<0.01.*

Figure 3-8: Example of Grand Marquis specifications

Car Specification	Pre-SID	Post-SID
Width (in)	77.8	78.2
Weight (lb)	3796	3958
Length (in)	212	212
Height (in)	56.8	56.8
Liter	4.6	4.6
Number of Doors	4	4
Model Year	1996	2000

Chapter 4

Nothing Happens in a Vacuum: Can a Dominant Metric be Shifted?

Abstract

Metrics and measurement standards play an important role in shaping the pace and direction of innovation. Yet in some cases, the dominant metric an industry orients itself around may lose, over time, some of its original ability to advance a field, or even introduce negative externalities. This paper studies whether it is possible to move firms away from a metric that has become a key focusing device for R&D and competition within their industry, and channel investments in R&D towards a new policy direction. To address this question, the paper takes advantage of a policy shock to estimate the effects of the "removal" of a metric (watts) within the European vacuum cleaner industry. While the objective of policymakers was to shift firms away from focusing on watts and improve energy efficiency, results show that there was no discernible change in the latter. Instead, the paper finds a severe crowding out effect, with firms reducing their R&D investment in the focal area affected by the policy. This relationship is larger in magnitude when accounting for patent quality, suggesting that this reduction is not simply driven by a lower number of fringe patents. Moreover, firms are not simply stopping their R&D investments but rather moving their effort towards adjacent product categories that are still unregulated. Taken together, the paper introduces the idea that removing a dominant metric in a given industry is not sufficient, possibly because of strong path dependency, to induce meaningful investments in a novel, desired direction. Combining the removal of a metric with additional support for lowering the costs of adopting an alternative metric is more likely to lead to meaningful, long term changes in innovation.

4.1 Introduction

In 1912, a pharmaceutical chemist in Detroit was working on a challenge: a popular muscle salve needed to scale in production, but some batches causing burns. His solution was to propose a novel measurement approach to standardize the amount of capsicum content in the salve formulation (Scoville, 1912). This chemist was Wilbur Scoville, and his Scoville Heat Units (SHU) became the dominant metric for gauging chili pepper pungency. SHU diffused widely within pharmacology as well as other domains such as agriculture, food sciences and culinary arts (Gmyrek, 2013). Despite its issues of human subjectivity and sensory sensitivity,[1] SHU evolved to become not only a unit for assessing existing peppers, but also a key target in the development of new pepper varietals. Furthermore, although alternative, more accurate measures of capsicum content have been developed (e.g. high performance liquid chromatography), SHU has persisted through the years, continuing to be the predominant approach for pepper pungency (Collins et al., 1995). This SHU example highlights that metrics can improve both understanding and execution of scientific innovation. Yet at the same time, there is a path dependent nature of measurement, in which a 'dominant' metric can persist in an industry for sustained periods of time (Anderson & Tushman, 1990; Utterback, 1994). Furthermore, in some cases, the metric may misdirect attention and effort towards continually improving upon it, potentially generating negative externalities (Kerr, 1975; Espeland & Sauder, 2007; Sauder & Espeland, 2009). This raises the question of whether it is possible to move

[1] SHU is measured through a ortholeptic test in which the pepper extracts are diluted until a panel of experts no longer taste spiciness.

innovation away from a metric that has become a key focusing device for R&D within an area?

This paper explores this question by estimating the impact of a unique policy intervention intended to shift an industry away from its dominant metric. In the early 2010s, European Union policymakers examined the state of domestic energy consumption and made a perceptive observation: while the watt was heavily used within the domestic vacuum cleaner industry to define performance, the actual ability of a vacuum cleaner to clean dust was poorly approximated by this de facto measurement standard. Moreover, in theory, the industry had access to alternative, less energy-consuming metrics (such as airflow and air watts) that would have been better able to capture cleaning performance. This led to a policy intervention targeted at eliminating any incentives for improving on watts by setting a cap on the input power that a vacuum cleaner manufactured or traded in Europe could have. Importantly, this was a meaningfully low cap that surprised those in the industry: at the time the cap was proposed, over 40% of the vacuum cleaners on the market exceeded the cap. Moreover, the regulators planned to introduce a further 44% reduction in the cap within 3 years from the first policy change. This regulatory intervention however, did not fully account for the dual role that the watts metric served in the industry: (1) on the one hand, it was driving firms towards producing end products with higher energy consumption, but (2) on the other hand, it was also a figure of merit that the industry had organized its innovation strategies and invested its R&D capabilities around. As a result, while the policy focused on addressing the former, it was not clear whether it could have additional side effects on the long run innovative behavior of the firms involved. In particular, because the cap essentially removed the ability to rely on watts to differentiate their products, it left industry participants in a state where they either had to search for and adopt a new metric, or otherwise adapt their product strategy to an alternative source of differentiation in the changed market conditions.

Using the universe of domestic cleaning granted patent applications (2005-2019), the paper shows that while the policy did reduce the wattage of end product vacuum cleaners available on the market, this did not translate into more firm innovation using energy efficient, alternative metrics. In other words, there is no discernible difference in the number of efficiency-related patents before versus after the shock. Instead, the policy is followed by a sharp drop in the number of vacuum cleaner-related patents, and this result is driven by firms that heavily rely on metrics in their patent applications (as defined before the policy is introduced). This relationship is larger in magnitude when accounting for patent quality, suggesting that this reduction is not simply driven by a lower number of fringe patents. Moreover, firms exhibit a shift from patenting in vacuum cleaners (the category affected by the policy) towards unregulated non-vacuum floor cleaning categories. This evidence suggests that rather than investing in search for and adapting to a new metric to organize innovation within vacuum cleaners, the affected firms sought out related areas where they could still make progress along the dominant metric they were accustomed to.

Taken together, the paper suggests that while removing a dominant metric within an industry may induce changes in the end product market, this is mainly out of compliance. This type of cap-based policy does not necessarily shift innovative behavior in the desired direction in the long term. Instead, firms can avoid searching for an alternative metric to orient innovation around by simply switching to an adjacent domain where they are able to retain the existing metric and continue innovating within their familiar structures. Policymakers or other principals seeking to induce an actual shift in innovation must therefore account for the costs of these two choices to firms, and identify ways in which the former can be made more attractive than the latter.

4.2 Related Literature

This paper relates to the literature on standards and innovation. Specifically, dominant metrics are cognate with de facto standards as they are non-rivalrous goods that arise without an explicit originator or sponsoring agency (David & Greenstein, 1990). Despite substantial theoretical attention to the determinants of these standards —such as path dependence (David, 1985), increasing returns to adoption (Farrell & Saloner, 1985; Katz & Shapiro, 1994), inertia (Katz & Shapiro, 1986), and invisibility in downstream markets (Bowker & Star, 1999; Timmermans & Epstein, 2010) —there has been relatively little empirical attention on whether and how policymakers or managers might endeavor to change them once ensconced.[2] Assuming the conditions the literature identifies exist, how can they be ameliorated? If the dominant metric is obsolete or a suboptimal standard, can simple simple external

[2] While this paper focuses on an informal measurement standard, the findings in this article complement that of Blind et al. (2017) which finds that in cases with uncertainty, regulations can have a negative impact on innovation relative to formal standards; as well as Simcoe (2012) which finds that politicization and distributional conflicts can interfere with standards development.

interventions be leveraged to shift the field away from it?

This paper also relates to the commensuration and reactivity literature, where the introduction of measurement can distort incentives, induce strategic behavior, and eliminate diversity (Sauder & Espeland, 2009; Mazmanian & Beckman, 2018; Espeland et al., 2016; Berman & Hirschman, 2018; Muller & Muller, 2018). Introducing quantification to a domain can cause the individuals or organizations that constitute the measurand to react in response to the metric itself, distorting the information that measurement can provide. As Espeland & Sauder (2007) find for instance, in the setting of law school rankings, the reactive behavior to progress along the metric led to persistent, reinforcing changes in who would get admitted and what was prioritized in teaching, shaping what law schools defined themselves as. Relatively less explored in this literature are later-stage settings when a quantitative metric is already in place, and participants within the domain have already invested time, resources and in some cases, their identities around the metric. Exploring settings with a long-term, dominant metric can shed light on how the consequences of commensuration might weigh against the stability in competition dynamics and identity that it creates. It also allows for the exploration of whether there is heterogeneity in how individuals or organizations respond towards efforts to move away from it.

More broadly, this paper contributes to our understanding of the determinants of the direction of scientific and technological innovation (reviewed in Cohen (2010) and Di Stefano et al. (2012)). Dominant metrics can moderate technological opportunity, influencing an organization's ability to understand, assimilate and exploit new information (Cohen & Levinthal, 1989). Dominant metrics can also be interpreted as a microfoundation of incumbent inertia, leading to rigidity in how organizations approach research and development (Leonard-Barton, 1992; Nelson, 1985; Gilbert, 2006).

4.3 Setting

4.3.1 Vacuum Cleaner Technology

Since the first horse-drawn vacuum cleaners trotted through the streets of England at the turn of the century, they attracted great interest and demand for automating a laborious and pervasive domestic cleaning task. Since that time, this demand has fueled countless iterations in product design and distribution, leading to it being along with refrigerators and microwaves, one of the most commonly owned home appliances. For instance, the stock of vacuum cleaners in EU households was 279.5 million in 2010, or roughly 2 vacuums for every 3 people, with annual sales in the range of 35 million new vacuum cleaners per year (Viegand Maagoe (2018)). Similarly, in the United States the stock was 113.4 million (roughly 1 vacuum for every 3 people), with annual sales of 26 million units (Energy EPA (2011)).

As a basic primer, modern consumer vacuum cleaners can generally be classified as (1) corded (in an upright or cylinder form factor) or (2) cordless (in an upright and/or handheld, and more recently, robotic form factor). At the basic level, vacuum cleaners consist of an intake nozzle, an electric motor and fan, a dust receptacle (bagged or bagless) and an exhaust port. Air and debris is pulled in through the intake nozzle by the suction pressure created by the motor and fan as it spins. As this passes through the dust receptacle, the debris is contained and clean air then continues to be pulled through, exiting through a filtered exhaust port.[3] The performance of a vacuum cleaner, in terms of how effectively it is able to clean particulate debris from a surface, therefore is a function of multiple factors and can be gauged by different metrics: the input power of the motor (e.g. watts is common to Europe, amps is more frequently used in the US, horsepower and joules are sometimes used); airflow (e.g. cubic feet per minute, air watts, metres per second); suction (e.g. water lift); bag filtration efficiency (e.g. particulate micron size). [45] In addition, there is a measure of dust pick-up performance which can be derived from running evaluations of vacuum cleaners in a controlled environment with a set mass of artificial particulates. Dust pick-up is computed as the ratio of the mass of particulates picked up in the vacuum cleaner relative to the total particulates. For this metric to be used beyond small samples however requires within-firm or industry standards on the characteristics of the test environment.[6]

While the exact path is unclear, watts emerged as a key metric upon which vacuum cleaners were and largely continue to be compared against one another despite the fact that it represents only the amount of power input that goes into the vacuum cleaner motor.[7] While the average vacuum cleaner had an input power of 500W in

[3] https://news.samsung.com/global/a-look-inside-your-vacuum-how-it-works

[4] In addition is portability (e.g. weight), maneuverability (e.g. nozzle width, hose and cable length), and noise (e.g. decibels) that are more secondary factors in overall vacuum cleaner performance.

[5] https://www.ristenbatt.com/xcart/Vacuum-Cleaner-Performance-Aspects.html

[6] See International Electrotechnical Commission 60312 for further details

[7] One possible channel is that the earliest versions of vacuum cleaners relied on steam engines for which horsepower is the standard performance metric. Another channel is that vacuum cleaner marketers amplified a “marketing war” centered on input power that fueled the emphasis on improvements along watts.

the 1960s, this had increased to 2000W in the 2010s (Reisch et al. (2010)). Moreover, this pace of watts improvements followed an upwards trend: a study of retail catalogues found that within only a 5 year period from 2003 to 2008, the maximum wattage increased by 200W for bagged upright and 700W for cylinder vacuums (Reisch et al. (2010)). Emphasis on watts was strong enough such that European advertising regulators clarified in 2010 that "*the product with the highest input watt motor may not have the best suction power. 'Power' claims that refer to a product's motor size e.g. 'powerful 1500W suction', should not be based on input wattage, because consumers are unlikely to interpret such claims in terms of power input only* (CAP (2010)).[8]

4.3.2 Vacuum Cleaner Industry

Although the vacuum cleaner industry originated at the turn of the century from firms solely focused on producing vacuums —for instance, Kirby, The Hoover Company, The Eureka Company initially only produced household upright and canister vacuum cleaners —as the industry matured, diversification became the strategic norm especially after World War II as household appliance purchases increased (Gantz, 2012). This has continued to this day where the modern vacuum cleaner industry is made up of firms that produce other home appliances, most commonly those with technological overlap such as floor care devices or implements for different surfaces (e.g. carpets, linoleum), air purifiers, and small motor-powered appliances

[8] The US EnergyStar had a similar analysis that "*Many [vacuum cleaner] brands feature amperage rating prominently on product packaging and associate it with cleaning performance prompting consumers to pick models that have the highest amperage rating. Wattage ratings are more frequently used in European markets.*" See EPA (2011) for more details.

such as hair dryers, blenders, and food processors.

There is a mixture of firms types including original equipment manufacturers (OEMs) that design, manufactures and assembles their products and components in facilities under their ownership (e.g. Miele fully owns its vacuum cleaner manufacturing plants in Germany, China, and the Czech Republic[9]), OEMs that subcontract with contract manufacturers (e.g. Bissell and others design their products and then contract with Flextronics, a leading contract manufacturer for floor care products[10]), and ODMs which design and develop vacuum cleaners and sell them as white labeled products to marketing distributors (e.g. KingClean). The vacuum cleaner industry has seen several contract manufacturers transition into becoming ODMs, joint ventures and some OEMs as well (e.g. LEXY). Intellectual property is generally developed and owned by OEMs and ODMs.

4.3.3 European Union policy

Motivated by a desire to decrease energy consumption within domestic appliances and consumer electronics, the European Union created the Ecodesign and Energy Labelling Regulations task force for proposing new regulations to encourage more energy-efficient product designs and information disclosures to consumers, and more broadly, reduce Europe's household energy consumption. Specific to vacuum cleaners in 2013, the task force passed directive No 666/2013 (commonly referred to as the "Ecodesign power limit") which set limits on the number of watts a vacuum cleaner

[9] See a full description of each factory here: https://www.miele.com/en/com/production-sites-2157.htm

[10] https://origin-sc.flex.com/industries/consumer-home-and-lifestyle/floor-care

manufactured, traded and sold in the EU could employ for its input power: within 1 year, the maximum power would be set at 1600 watts; and within 3 years, it would be set at 900 watts.[11]

While improvements in the environmental-friendliness of domestic appliances could have been anticipated,[12] the timing and the magnitude of the decrease was not. In particular the maximum wattage threshold was a significant reduction from the industry norms. When this directive was proposed, roughly 40% of vacuum cleaners on the market exceeded 1800W. This spurred lobbying from vacuum cleaner firms, and negative media coverage in affected European countries.

A subsequent report commissioned by the EU to evaluate the policy found that firms did comply and wattage decreased accordingly in vacuum cleaners on the market: "*in only a few years, the Ecodesign and Energy Labelling Regulations have revolutionised the vacuum cleaner market...it is seen from the available market data that the energy consumption of all regulated types of vacuum cleaners have decreased around 40% from introduction of the Ecodesign and Energy Labelling Regulations in 2013 to 2016.*" (Viegand Maagoe (2018)). The policy evaluation relied exclusively on retail market data to compare the average input power on the market against projected counterfactuals. This was then translated into overall European vacuum cleaner energy consumption, finding an overall decline from 78 kWh/year in 2010

[11] There was also a cap on the noise level of the vacuum in terms of decibles emitted in operation, however this was labeled a "second tier" requirement and was not significantly lower than industry offerings at the time. In addition, there was additional non-metric requirements for a set minimum level for dust pick up, the vacuum hose durability and operational motor life time (Viegand Maagoe (2018)).

[12] Given the 2007 European Union commitment to reduce domestic carbon emissions by at least 20% by 2020, and the passing of similar regulations with maximum wattage caps in incandescent light bulbs and standby power wattage for plasma televisions

to 34 kWh/year in 2016 for domestic cylinder vacuum cleaners. It bears mention that the derivation of energy consumption relies on a constant set of assumptions including on a more macro level (e.g. constant purchasing behavior in terms of unit replacement, average number of vacuums purchased per household); and more micro such as the number of strokes vacuumed over a surface (2 double strokes) and number of cleaning events per year (50 cycles per year). This suggests that any changes in consumer vacuuming behavior for lower wattage vacuums are not accounted for (e.g. increasing strokes out of perceived need to do so even if cleaning effectiveness is unchanged).

The official policy evaluation report centers on the end products available in the retail market for consumers. This is consistent with the mission to reduce domestic energy consumption in Europe, but it does not consider the broader impact of curbing a dominant metric on firms' innovation trajectories. The focus of this study is to fill that void, and explore how the policy impacted not just end products, but more upstream innovation as well.

4.4 Data and Methods

This section describes the data and sample construction used to examine the impact of the metric cap on firm innovation between 2005 and 2019. To unpack how firms respond, the analysis is conducted the firm-year level, with treatment determined by the concentration of patenting activity prior to the policy intervention. This section describes the main data set used, outlining the approach to treatment assignment and variable construction.

4.4.1 Patent Applications Data

The primary data set used for this study comes from the full corpus of patent applications that were filed with the US Patent and Trademark Office (USPTO) between 2005 and 2019. These are applications that eventually were granted a patent, and capture innovations that have reached a level where they could be produced or commercialized as a corporate invention or product.

The USPTO uses the Cooperative Patent Classification (CPC) Scheme to assign patents to the most relevant technology class(es). The CPC taxonomy is made of 5 levels consisting of a section (e.g. A -" Human Necessities"), class (A47- "Furniture; Domestic Articles or Appliances; Coffee Mills; Spice Mills; Suction Cleaners in General"), subclass (A47L-"Domestic Washing and Cleaning"), main group (A47L5- "Structural features of vacuum cleaners"), and subgroup (A47L5/02-"User-driven air pumps or compressors"). The main sample focuses on the 533 subgroups within the subclass A47L "Domestic Washing or Cleaning; Suction Cleaners in General" which covers machines or implements used for household cleaning. Patents were assigned based on their primary subgroup classification. The subgroup definitions were used to classify patents as being related to domestic upright and canister vacuum cleaners. These patent subgroups are referred to as "domestic vacuums" in this paper and cover innovations that are treated by the policy. Two examples are A47L9/2831 - "Motor parameters e.g. motor load or speed" and A47L9/2842 - "Suction motors or blowers."

Subgroups were also used to further classify patents into two additional categories. This included those related to floor care that were not domestic upright or canister vacuum cleaners such as A47L 7/0004 - “Suction cleaners adapted to take up liquids, e.g. wet or dry vacuum cleaners” and A47L11/28 - “Floor-scrubbing machines, motor-driven”. Patents were also classified into domestic cleaning that did not involve floor care or vacuum cleaners (e.g. A47L 1/05-“Cleaning windows, Hand apparatus with built-in electric motors”). The full list of subgroups classified into each of these categories is in the Appendix.

4.4.2 Treatment Assignment

Given that firms in the vacuum industry largely hold innovation portfolios of several different products, treatment assignment was identified based on their ex ante concentration of innovative effort. To measure this, patent applications filed in the years preceding the policy were collected for all firms. This was used to identify the subgroups that each firm was patenting in prior to the policy. The average firm in this period patented in 3.95 different subgroups. Firms that patented the most frequently in domestic vacuums prior to the cap were identified as treated by the policy. These are firms that had heavily invested in innovation in domestic vacuum cleaners before the unexpectedly strict cap was announced. The main control group in the analysis are firms where the most frequently patented subgroup prior to the cap was within A47L, but was not in domestic vacuum cleaners. These are firms with innovation capabilities that are related to domestic vacuums, but were relatively less affected by the cap as it did not affect their area of focus. To make this concrete with examples following this approach, Bissell, Miele and the Hoover Company are firms that are assigned to the treated group as domestic vacuums make up their highest share of prior patents. Kaercher, Bosch and Rug Doctor are firms assigned to the control group as they patent in related subgroups, but allocated relatively less of their ex ante innovative effort to domestic vacuums.

4.4.3 Variable Construction

To understand more about the quality and content of the patents in the applications data, patent citations and summary text for each patent were collected. In order to gauge firms' responses against the policymakers intentions, the text data was analyzed to determine the extent to which a patent involved the use of watts (or a cognate metric), as well as the extent to which it contributed to energy efficiency. These classifications of the patents then allowed for the firms' orientation towards metrics, watts and energy efficiency in their innovation portfolios.

The patent summary text was then processed using the text classifier Quantulum to parse and identify all written and numeric quantities (e.g. '20' or 'twenty'). This also allowed for the extraction of the associated metric (e.g. 'watt' or 'W') and the broader entity the metric is linked to on Wikipedia (e.g. 'watt' and 'ampere' both belong to the entity 'power'.) To hone in on scientific and technical metrics related to innovation, time, count, percentage, and financial metrics were excluded. Together, this analysis of the patent summary text was used to create two variables. First, a simple indicator variable was used to classify whether the patent had an above the median number of uses of metrics in its summary text. Second, an indicator was used to classify whether the patent directly referenced watts (or its associated plural and unit representations), or a related power measure in the summary text (e.g. volt, ampere, horsepower or joule). Finally, the summary text was used to measure the number of keywords related to improvements in energy efficiency.[13] This was used to code a dummy indicating whether a patent had above the median energy efficiency related text. Together, the summary text provided an indication of whether a patent was measurement-focused, involved the use of watts, and whether it was related to improvements in energy efficiency.

The classification of patents based on their metric content was also used to determine if firms had invested in developing metrics-driven innovation capabilities or not. An variable for whether a firm is a publicly traded firm covered in Compustat was created to proxy for firm size and resources following the approach in Autor et al. (2020b).

4.4.4 Summary Statistics

Together, the sample used in the analysis is a panel spanning 2005 to 2019 in patent application years (Table 1). There are 492 firms observed (132 treated) resulting in 7,380 firm-year observations. Of the 492 firms in the sample, about 41% were producing relatively more metrics-focused patents, and 9% were public or economically significant private firms that were covered by Compustat. The average firm applied for 6.4 patents per year (that eventually were granted), corresponding to 24% patents that were related to domestic vacuums, 16% patents related to non-vacuum

[13] This was a straightforward search for the the word stems of improved and efficiency, along with synonyms; as well as ratio metrics related to the policy intention such as cubic feet per metre.

floor care per year, and the remainder towards other domestic cleaning.

4.4.5 Empirical Strategy

The analysis relies on a standard difference-in-differences approach to estimate the impact of the policy on firm innovation:

$$\text{Patents}_{i,t} = \alpha + \beta \text{Treated}_i \times \text{Post}_t + \lambda_i + \delta_t + \epsilon_{i,t} \tag{4.1}$$

where Patents_{it} captures the patent innovation activity from firm i in year t. Given the skewed nature of patent output, the dependent variables are log transformations of patent counts. To understand whether innovation could be shifted from a dominant metric through a cap, the analysis examined the impact on three types of innovation outcomes: energy efficiency-related patents, domestic vacuum patents, and watts-related patents. As discussed above, Treated_i represents the treated group of firms that are expected to respond to the cap on watts. The dummy Post_t equals 1 for each year after the passing of the policy in 2013. As such, β is the main coefficient of interest. λ_i and δ_t are firm and year fixed effects respectively. Standard errors are clustered at the firm level in all regressions.

4.5 Results

4.5.1 Energy Efficiency Innovation

As a first step of the analysis, the data was used to understand whether the policy achieved its intended objective of increasing investments in energy efficiency-related innovation. Table 2 presents the difference-in-differences estimates in equation 1 where Patents_{it} is the logged count of energy efficiency-related (herein EE) patents from firm i in year t. Column 1 begins with examining any changes to EE patenting within domestic vacuum cleaner patent subgroups. Column 2 follows by expanding this to examine changes to EE patenting in all domestic cleaning patent subgroups, including vacuum cleaners, other floor care, and other household cleaning. Column 3 takes into account firms' full patenting activity. Across these three levels, there is no discernible change to energy-efficiency related patenting. This suggests that in the short term (within 6 years) following the introduction of the cap, the policy did not induce the desired shift in innovation effort towards energy efficiency among domestic vacuum firms, relative to the control group of non-vacuum household cleaning firms. Column 4 presents an estimate of the effect on overall patenting and finds no statistically significant change. Together, these estimates suggest that firms did not shift on the extensive margin towards energy efficiency innovation, nor did they change their level of overall patenting. This raises the question of whether there was a discernible firm response to the cap in terms of innovation.

4.5.2 Domestic Vacuum Cleaners

To examine this question, the analysis hones in on domestic vacuum cleaning patents. Column 1 of Table 3 presents estimates of equation 1 where Patents_{it} is the logged count of domestic vacuum cleaner patents from firm i in year t. It shows that after the introduction of the cap in 2013, domestic vacuum cleaner patents in treated firms experienced an average decrease of 7.8% per year relative to control firms. Column 2 presents an estimate that is similar in magnitude and direction (-9.9%) when the dependent variable is transformed using the inverse hyperbolic sine function rather than logged (allowing it to retain zeroes). To understand if this decline is driven by low-impact, excess patenting, Column 3 employs a proxy for patent quality using patent citations. The observed decline of 20.3% in citation-weighted patents provides confidence in the interpretation that patenting activity in domestic vacuums underwent a meaningful decline after the policy.

To unpack the year-specific differences between treatment and control firms after the cap, the following equation was estimated:

$$\text{Patents}_{i,t} = \alpha + \sum_{t} \beta_t \text{Treated}_i \times \text{Year}_t + \lambda_i + \delta_t + \epsilon_{i,t} \tag{4.2}$$

where the treated group is interacted with a set of indicator variables, Year_t corresponding to a particular year relative to the introduction of the policy in 2013.

Figure 4-1 presents the estimates of β_t in a graph, along their 95-percent confidence intervals. Before the cap, the estimated coefficients are not statistically different from those in 2013. After, the estimate is lower in the two years following the cap (than the preceding period), with the magnitude of the decline increasing and becoming statistically significant after two years. This suggests that there was a lag period between the policy and the firms' innovation response. This might have happened if firms had some patents in the pipeline and took some time to adjust. Running the baseline regression in Table 3 dropping the observations in 2013, 2014 and 2015 finds a large and significant decline of 12.1% in domestic vacuum patents (-24.4% in citation-weighted) in the years 2016 and after relative to the pre-period.

To understand if this decline was observed across all firms, or if there were heterogeneous firm responses, the sample was split into firms that had ex ante higher level of measurement incorporated into their innovation approaches. Firms that have a metrics-orientation may have specific incentive schemes, equipment, and human capital for making progress along the dominant metric, making it more difficult for them to engage in exploratory activities or pivot towards a different direction. Table 4 presents the estimates from this approach, essentially splitting Column 1 and 3 of Table 3 into firms that were relatively more measurement-oriented versus less so prior to the policy. Column 1 shows that metric-oriented firms reduced their domestic vacuums patents by 10.7%. In contrast, firms that were less oriented around measurement experienced a smaller, and statistically insignificant decline of 5.2%. Using logged citation-weighted patents as the dependent variable in columns 3 and 4 finds a consistent pattern in which metrics-oriented firms see a larger decline in domestic vacuums. This suggests it is not simply low-quality firms that drop out because of the costs of switching from watts, but meaningful innovation contributors that are affected as well. Together, this suggests that the cap on watts reduced domestic vacuum cleaner patents for all firms, but had a greater effect on those that demonstrated a greater ex ante emphasis on measurement in their patenting approach.

Table 5 replicates the above analysis with a more narrowly defined dependent variable as domestic vacuum patents that explicitly mention watts or a related power metric in their patent summary text. Column 1 replicates Table 3 Column 1, finding a 3% reduction in watts-related domestic vacuum patents. Columns 2 and 3 replicate Table 4 Columns 1 and 2 respectively, similarly finding that there is a decline in patents among all treated firms, with a higher decrease among metric-oriented firms (-4.3%) versus less measurement focused firms (-1.9%). The estimates are smaller in magnitude, which is in line with the use of the more conservative variable, but follow the general pattern of a decrease that is driven mainly by metric-oriented firms. The overall stability of the observed decline in patents increases the confidence in the interpretation of these estimates as responses to the cap.

4.5.3 Alternative Adjacent Areas

The previous sections find that there is a decline in domestic vacuum cleaner patents, yet at the same time, overall patent rates remain stable. This suggests that firms are not simply pausing and reassessing their innovation strategies, but rather potentially substituting their efforts in a different direction. For instance, a low cost way to respond to the policy would be to continue the extant innovation approach with the dominant metric in an adjacent, unregulated area.

To examine this, equation 1 was estimated with a focus on understanding the impact on innovation in in floor care excluding domestic upright and canister vacuum cleaners. Floor care technologies and products such as wet dry vacuums, floor scrubbers and buffers, handheld vacuums, steam mops, and sweepers constitute an adjacent area of innovation, with significant overlap in many facets of design and engineering but are not affected by the cap on watts. Table 6 presents the findings from this analysis. Column 1 finds that there is a positive, statistically significant increase in (non-vacuum) floor care patents by 2.9% from treated firms after the policy intervention. Column 2 finds that the estimate is stable when using the inverse hyperbolic sine function instead of a log transformation. Column 3 shows that there is a positive but statistically insignificant increase when patents are weighted by citations.

Finally, to explore whether these are the same types of firms that are driving this switch towards unregulated, adjacent floor cleaning, Table 7 splits the sample into metric-oriented firms versus non-metric firms (following the approach in Table 4). The estimates in Columns 1 and 2 show that as with the decrease in vacuum cleaner patents, it is primarily metric-focused firms that are driving the increase. Columns 3 and 4 show that metric-oriented firms produce high impact, high quality patents in the floor care space (+7.9%) relative to those that rely less explicitly on measurement.

To summarize the overall results, after the cap on watts was introduced for domestic vacuum cleaners in 2013: (1) innovation in energy efficiency remained unchanged; (2) firms decreased innovation both in terms of quantity and quality in the target area of the policy: domestic vacuum cleaners; and (3) firms shifted innovation towards the adjacent area of floor care that was unaffected by the policy. Taken together, this suggests that setting a cap on a dominant metric does not engender exploration of new, alternative metrics. It is possible that it is too soon to observe a change in innovation towards energy efficiency and/or for a new 'figure of merit' to emerge in the vacuum cleaner industry. With this caveat in mind, in the short term, the entrenched metric can continue to shape innovation as firms find ways to circumvent the constraints introduced by the policy.

4.6 Conclusion

Convergence on a measurement can determine how innovation is produced, rewarded and defined. At the same time, these dominant metrics can, possibly because of incentives and path dependency, lead to settings in which firms and individuals become entrenched in focusing their attention and resources on the metric. This can be of concern when the metric becomes obsolete or is found to create negative externalities. This raises the question of whether it is possible to redirect innovative effort away from the dominant metric?

This paper studies this question by examining the impact of a simple and unexpected policy intervention, the use of a cap on the prevailing metric of watts, on domestic vacuum cleaners in 2013. This paper finds that shifting the dominant metric using such an approach is not straightforward, and a cap may not induce meaningful change. Within the domestic vacuum cleaner industry, the cap on led to no significant change in innovation towards energy efficiency improvements, which was the intention of the policy. Instead, the cap significantly decreased patenting activity within domestic vacuum cleaners by 7.8-9.9% (and a 20% decrease when considering patent quality). This had a larger effect on firms that placed a greater emphasis on measurement in their innovation approach prior to the cap. Moreover, without a clear incentive to bear the costs of adopting an alternative metric, these types of interventions may lead firms towards strategic behaviors. A policy that only targets one area may simply shift innovative effort to continue advancement along the dominant metric in adjacent and unregulated areas. This paper finds evidence in support of such a channel, with firms increasing their innovative output in the unregulated area of non-vacuum cleaner floor care.

Overall, this paper introduces the idea that dominant metrics can be entrenched in an industry. While in principle, reducing the rewards to advancing along a dominant metric could lead firms to reducing effort in that realm, considering the costs of exploring a new domain makes the reallocation of innovative effort more ambiguous. In settings such as this one, in which a metric has become a key dimension of strategic competition, removing the metric through an external cap did reduce effort in the targeted area but rather than shifting it in line with the intentions of the policy, it led to skirting of the cap.

While this paper examines only one policy lever, it suggests that policymakers or managers seeking to displace a dominant metric should carefully consider the scope of any interventions they introduce. Because dominant metrics are often the main frame that innovators use to define progress, even in the absence of pecuniary incentives, there can be inertia in moving away from it. Potential strategies that can be explored include combining a cap with concerted efforts to lower the costs of exploring alternative metrics, for example through research, forums for collaboration across firms, and funding for managing the transition. Another approach would be to forgo the cap, and instead, increase the incentive to work on alternative metrics by mandating their disclosure. This effectively elevates a metric to a dominant one through external intervention. At the time of writing, EU policymakers seem to be moving in this direction and is in the process of passing a standardized label with multiple metrics related to vacuum cleaner energy efficiency that must be disclosed. If passed, future research on whether this ameliorates the effects of the cap could be a worthwhile endeavor. This paper leaves several other open question for future research. Because this paper explores an understudied facet in prior work of exploring metrics as a type of informal standard, it would be valuable to understand whether the findings hold in less mature industries where firms may be more experimental. In addition, understanding whether consumers become similarly entrenched in using dominant metrics in their valuation of products could shed light on the mechanisms behind policy circumvention. Ultimately, deepening our understanding on dominant metrics in firms is fruitful as it can allow for active management of this driver of innovation.

Tables

Table 1: Summary Statistics

Variable	Mean	Std. Dev.	Min.	Max.	N
Patent Count	6.392	42.199	0	1,185	7,380
Domestic Vacuum Patents	0.236	1.796	0	58	7,380
Floor Cleaning Patents	0.158	0.825	0	19	7,380
Treated	0.254	0.435	0	1	7,380
Post	0.4	0.49	0	1	7,380
Year	2012	4.321	2005	2019	7,380
Metric Firm	0.413	0.492	0	1	7,380
Large Firm	0.09	0.286	0	1	7,380

This table provides information on the main sample. The unit of analysis is the firm-year level. The data source is the USPTO PatentsView dataset on patent applications in the CPC subclass A47L (Domestic Washing or Cleaning; Suction Cleaners in General) for years 2005-2019. Vacuum and Floor Cleaning patents are assigned based on the subgroup definitions. The treated group is defined as firms who patented primarily in vacuum cleaners before the policy intervention. This is defined as A47L being among the most frequent subclass the firm patented in. Post-treatment is defined as the years after 2013. Metric firms are defined as those that patent an above-median amount of metric-focused patents ex-ante. Metric-focused patents are those for whom their patent summary referenced an above-median number of scientific or technical metrics. Large firms are defined as those that are publicly listed and covered by Compustat.

Table 2: Energy efficiency-related patents do not see change

	(1)	(2)	(3)	(4)
DV:	Log(EE Vacuums)	Log(EE Floor Care)	Log(EE All)	Log(All Patents)
Post x Treated	-0.005	-0.004	-0.011	-0.003
	(0.004)	(0.004)	(0.020)	(0.041)
Firm FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
N	7,380	7,380	7,380	7,380

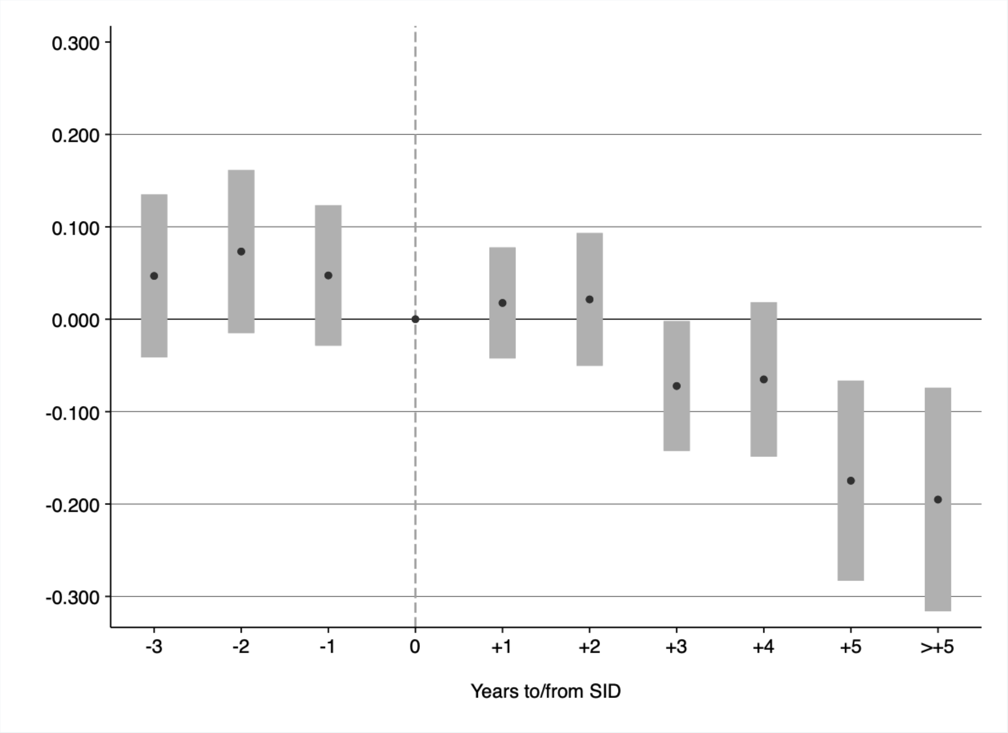
This table shows the relationship between being treated by the policy and energy efficiency-related patents. The dependent variable is the logged count of vacuum patents in that are energy efficiency related in domestic vacuum cleaners (Column 1), floor care (Column 2) and domestic cleaning overall (Column 3). Column 4 examines the impact on the logged count of overall patents. Post refers to years after the policy was passed in 2013, and Treated are the firms that were patenting disproportionately ex ante in the affected patent subclasses. The equations are estimated using a fixed effects regression model. Standard errors clustered at the firm level are reported in parentheses. (*=p < 0.10, **=p < 0.05 ,***=p < 0.01)

Table 3: Vacuum patents decline after the removal of the metric

	(1)	(2)	(3)
	Log(Vacuum Patents)	asinh(Vac)	Log(CW Vac)
Post x Treated	-0.080***	-0.103***	-0.233***
	(0.024)	(0.030)	(0.055)
Firm FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
N	7,380	7,380	7,380

This table shows the relationship between being treated by the policy and vacuum cleaner patents. The dependent variable is a measure of innovation in vacuum cleaners and is defined as the logged count of vacuum patents, the inverse hyperbolic sine of patent count, and the logged citation-weighted patents in Columns 1, 2, and 3 respectively. The equations are estimated using a fixed effects regression model. Standard errors clustered at the firm level are reported in parentheses. (*=p < 0.10, **=p < 0.05 ,***=p < 0.01)

Figure 4-1: Effect of removing watts on patenting in vacuums



The dots in the above plots are the coefficient estimates from the regression of logged vacuum cleaner patent counts on interaction terms between being a treated firm times the distance in years from the policy along with year and firm fixed effects. The shaded bars are the 95 percent confidence intervals.

Table 4: Metric-focused firms driving the decline in vacuum patents

DV	(1) Log(Vacuum Patents) Metric	(2) Non-Metric	(3) Log(Cite-wgt Vac Pat) Metric	(4) Non-Metric
Post x Treated	-0.113***	-0.053	-0.276***	-0.197***
	(0.034)	(0.034)	(0.095)	(0.063)
Firm FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
N	3,045	4,335	3,045	4,335

This table shows the relationship between being treated by the policy and vacuum cleaner patents split by metric focused firms versus non-metric focused firms. A metric-focused firm is defined as having an above the median number of patents with scientific and technical metrics mentioned in its patent summary text ex ante of the policy. The equations are estimated using a fixed effects regression model. Standard errors clustered at the firm level are reported in parentheses. (*=p $<$ 0.10, **=p $<$ 0.05 ,***=p $<$ 0.01)

Table 5: Decline persists to only looking at watts-related vacuum patents

	(1)	(2)	(3)
DV:	Log(Watts-related Vacuum Patents)		
	All	Metric Firms	Non-Metric Firms
Post x Treated	-0.030***	-0.044***	-0.019*
	(0.009)	(0.015)	(0.011)
Firm FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
N	7,380	3,045	4,335

This table shows the relationship between being treated by the policy and watts-related vacuum cleaner patents. The dependent variable is a measure of metric-driven innovation in vacuum cleaners and is defined as the logged count of watts-related vacuum patents. Columns 2 and 3 split by metric focused firms versus non-metric focused firms respectively. A metric-focused firm is defined as having an above the median number of patents with scientific and technical metrics mentioned in its patent summary text ex ante of the policy. The equations are estimated using a fixed effects regression model. Standard errors clustered at the firm level are reported in parentheses. (*=$p < 0.10$, **=$p < 0.05$,***=$p < 0.01$)

Table 6: Shift in innovation towards related, unregulated floor cleaning category

	(1) Log(Floor Cleaning Patents)	(2) asinh(Floor Pat)	(3) Log(CW Floor Pat)
Post x Treated	0.029***	0.038***	0.028
	(0.009)	(0.012)	(0.026)
Firm FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
N	7,380	7,380	7,380

This table shows the relationship between being treated by the policy and non-vacuum, floor cleaning cleaner patents. The dependent variable is a measure of innovation and is defined as the logged count of floor cleaning patents, the inverse hyperbolic sine of patent count, and the logged citation-weighted patents in Columns 1, 2, and 3 respectively. The equations are estimated using a fixed effects regression model. Standard errors clustered at the firm level are reported in parentheses. (*=$p < 0.10$, **=$p < 0.05$,***=$p < 0.01$)

Table 7: Metric-specialized firms driving the shift

	(1)	(2)	(3)	(4)
DV	Log(Floor Cleaning Patents)		Log(Cite-wgt Floor Pat)	
	Metric	Non-Metric	Metric	Non-Metric
Post x Treated	0.044***	0.019	0.076***	-0.009
	(0.014)	(0.013)	(0.026)	(0.043)
Firm FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
N	3,045	4,335	3,045	4,335

This table shows the relationship between being treated by the policy and non-vacuum floor cleaning patents split by metric focused firms versus non-metric focused firms. A metric-focused firm is defined as having an above the median number of patents with scientific and technical metrics mentioned in its patent summary text ex ante of the policy. The equations are estimated using a fixed effects regression model. Standard errors clustered at the firm level are reported in parentheses. (*=p $<$ 0.10, **=p $<$ 0.05 ,***=p $<$ 0.01)

Bibliography

AGARD (1996). *Anthropomorphic Dummies for Crash and Escape System Testing*. Defense Technical Information Center.

Ahrens, A., Hansen, C. B., & Schaffer, M. E. (2020). lassopack: Model selection and prediction with regularized regression in stata. *The Stata Journal*, *20*(1), 176–235.

Almond, D., Doyle Jr, J. J., Kowalski, A. E., & Williams, H. (2010). Estimating marginal returns to medical care: Evidence from at-risk newborns. *The Quarterly Journal of Economics*, *125*(2), 591–634.

Anderson, P., & Tushman, M. L. (1990). Technological discontinuities and dominant designs: A cyclical model of technological change. *Administrative Science Quarterly*, (pp. 604–633).

Arora, A., & Gambardella, A. (1994). The changing technology of technological change: general and abstract knowledge and the division of innovative labour. *Research policy*, *23*(5), 523–532.

Arrow, K. J. (1974). *The Limits of Organization*. WW Norton and Company.

Autor, D., Dorn, D., Hanson, G. H., Pisano, G., & Shu, P. (2020a). Foreign competition and domestic innovation: Evidence from us patents. *American Economic Review: Insights*, *2*(3), 357–74.
URL `https://www.aeaweb.org/articles?id=10.1257/aeri.20180481`

Autor, D., Dorn, D., Hanson, G. H., Pisano, G., & Shu, P. (2020b). Foreign competition and domestic innovation: Evidence from us patents. *American Economic Review: Insights*, *2*(3), 357–74.
URL `https://www.aeaweb.org/articles?id=10.1257/aeri.20180481`

Azoulay, P., Graff Zivin, J. S., & Manso, G. (2011). Incentives and creativity: evidence from the academic life sciences. *The RAND Journal of Economics*, *42*(3), 527–554.

Baker, G. P., & Hubbard, T. N. (2003). Make versus buy in trucking: Asset ownership, job design, and information. *American Economic Review*, *93*(3), 551–572.

Basker, E., & Simcoe, T. (2021). Upstream, downstream: Diffusion and impacts of the universal product code. *Journal of Political Economy*, *129*(4), 1252–1286.

Belloni, A., Chernozhukov, V., & Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, *81*(2), 608–650.
URL `https://doi.org/10.1093/restud/rdt044`

Berman, E. P., & Hirschman, D. (2018). The sociology of quantification: Where are we now? *Contemporary Sociology*, *47*(3), 257–266.

Blind, K., Petersen, S. S., & Riillo, C. A. (2017). The impact of standards and regulation on innovation in uncertain markets. *Research Policy*, *46*(1), 249–264.

Booth, C. M., & Eisenhauer, E. A. (2012). Progression-free survival: meaningful or simply measurable? *Journal of Clinical Oncology*, *30*(10), 1030–1033.

Bowker, G., & Star, S. L. (1999). Sorting things out. *Classification and its consequences*, *4*.

Bresnahan, T. F., & Yao, D. A. (1985). The nonpecuniary costs of automobile emissions standards. *The RAND Journal of Economics*, (pp. 437–455).

Brunner, K., Meltzer, A., et al. (1983). Econometric policy evaluation: A critique. In *Theory, Policy, Institutions: Papers from the Carnegie-Rochester Conferences on Public Policy*, vol. 1, (p. 257). North Holland.

Burrell, J., & Fourcade, M. (2021). The society of algorithms. *Annual Review of Sociology*, *47*, 213–237.

CAP (2010). Vacuum cleaner marketing: Advertising guidance. Tech. rep., Committee of Advertising Practice.

Chagné, D., Dayatilake, D., Diack, R., Oliver, M., Ireland, H., Watson, A., Gardiner, S. E., Johnston, J. W., Schaffer, R. J., & Tustin, S. (2014). Genetic and environmental control of fruit maturation, dry matter and firmness in apple (malus× domestica borkh.). *Horticulture Research*, *1*.

Chatterji, A. K., & Toffel, M. W. (2010). How firms respond to being rated. *Strategic Management Journal*, *31*(9), 917–945.

Chen, S., Chernozhukov, V., Fernández-Val, I., & Luo, Y. (2019). Sortedeffects: Sorted causal effects in r.

Chernozhukov, V., Fernandez-Val, I., & Luo, Y. (2018). The sorted effects method: Discovering heterogeneous effects beyond their averages.

Choi, J., Hecht, G. W., & Tayler, W. B. (2012). Lost in translation: The effects of incentive compensation on strategy surrogation. *The Accounting Review*, *87*(4), 1135–1163.

Choi, J., Hecht, G. W., & Tayler, W. B. (2013). Strategy selection, surrogation, and strategic performance measurement systems. *Journal of Accounting Research*, *51*(1), 105–133.

Christensen, C. M. (1997). *The innovator's dilemma: when new technologies cause great firms to fail*. Harvard Business Review Press.

Christin, A. (2018). Counting clicks: Quantification and variation in web journalism in the united states and france. *American Journal of Sociology*, *123*(5), 1382–1415.

Christin, A., & Lewis, R. (2021). The drama of metrics: Status, spectacle, and resistance among youtube drama creators. *Social Media+ Society*, *7*(1), 2056305121999660.

Clark, K. B., Chew, W. B., Fujimoto, T., Meyer, J., & Scherer, F. (1987). Product development in the world auto industry. *Brookings Papers on Economic Activity*, *1987*(3), 729–781.

Cohen (2010). Fifty years of empirical studies of innovative activity and performance. *Handbook of the Economics of Innovation*, *1*, 129–213.

Cohen, & Levinthal (1990). Absorptive capacity: A new perspective on learning and innovation. *Administrative Science Quarterly*, (pp. 128–152).

Cohen, W. M., & Levinthal, D. A. (1989). Innovation and learning: the two faces of r & d. *The economic journal*, *99*(397), 569–596.

Collins, M. D., Wasmund, L. M., & Bosland, P. W. (1995). Improved method for quantifying capsaicinoids in capsicum using high-performance liquid chromatography. *HortScience*, *30*(1), 137–139.

Crosby, A. W. (1997). *The Measure of Reality: Quantification in Western Europe, 1250-1600*. Cambridge University Press.

David, & Greenstein (1990). The economics of compatibility standards: An introduction to recent research. *Economics of Innovation and New Technology*, *1*(1-2), 3–41.

David, P. A. (1985). Clio and the economics of qwerty. *The American Economic Review*, *75*(2), 332–337.

Dee, T. S. (1999). State alcohol policies, teen drinking and traffic fatalities. *Journal of Public Economics*, *72*(2), 289–315.

Di Stefano, G., Gambardella, A., & Verona, G. (2012). Technology push and demand pull perspectives in innovation studies: Current findings and future research directions. *Research Policy*, *41*(8), 1283–1295.

Dongarra, J. (2006). Trends in high-performance computing. In *Handbook of Nature-Inspired and Innovative Computing*, (pp. 511–520). Springer.

Dongarra, J., Moler, C. B., Bunch, J. R., & Stewart, G. W. (1979). *LINPACK users' guide*. SIAM.

Dosi, G. (1982). Technological paradigms and technological trajectories: a suggested interpretation of the determinants and directions of technical change. *Research policy*, *11*(3), 147–162.

Drucker, P. (1965). *The practice of management*. Routledge.

EPA, U. (2011). Energy star market & industry scoping report vacuum cleaners. Tech. rep., Energy Star.

Espeland (1993). Power, policy and paperwork: the bureaucratic representation of interests. *Qualitative Sociology*, *16*(3), 297–317.

Espeland, W. N., & Sauder, M. (2007). Rankings and reactivity: How public measures recreate social worlds. *American Journal of Sociology*, *113*(1), 1–40.

Espeland, W. N., Sauder, M., & Espeland, W. (2016). *Engines of anxiety: Academic rankings, reputation, and accountability*. Russell Sage Foundation.

Espeland, W. N., & Stevens, M. L. (2008). A sociology of quantification. *European Journal of Sociology/Archives Européennes de Sociologie*, *49*(3), 401–436.

Farrell, J., & Saloner, G. (1985). Standardization, compatibility, and innovation. *the RAND Journal of Economics*, (pp. 70–83).

Foucault, M. (1977). *Discipline and punish: The birth of the prison*. Vintage Books.

Fourcade, M. (2011). Cents and sensibility: Economic valuation and the nature of "nature". *American journal of sociology*, *116*(6), 1721–77.

Freyaldenhoven, S., Hansen, C., & Shapiro, J. M. (2019). Pre-event trends in the panel event-study design. *American Economic Review*, *109*(9), 3307–38.

Gans, J. S., Stern, S., & Wu, J. (2019). Foundations of entrepreneurial strategy. *Strategic Management Journal*, *40*(5), 736–756.

Gantz, C. (2012). *The vacuum cleaner: a history*. McFarland.

Gavetti, G., Levinthal, D. A., & Rivkin, J. W. (2005). Strategy making in novel and complex worlds: The power of analogy. *Strategic Management Journal*, *26*(8), 691–712.

Gibbons, R., LiCalzi, M., & Warglien, M. (2021). What situation is this? shared frames and collective performance. *Strategy Science*, *6*(2), 124–140.

Gibbons, S., Machin, S., & Silva, O. (2013). Valuing school quality using boundary discontinuities. *Journal of Urban Economics*, *75*, 15–28.

Gilbert, C. G. (2006). Change in the presence of residual fit: Can competing frames coexist? *Organization Science*, *17*(1), 150–167.

Giorgi, M., & Edward (2019). On the relationship between firms and their legal environment: the role of cultural consonance. *Organization Science*, *30*(4), 803–830.

Gmyrek, D. P. (2013). Wilbur lincoln scoville: the prince of peppers. *Pharmacy in History*, *55*(4), 136–156.

Goel, V. (1992). Comparison of well-structured & ill-structured task environments and problem spaces. In *Proceedings of the fourteenth annual conference of the cognitive science society*, (pp. 844–849). Citeseer.

Goke, A., Serra, S., & Musacchi, S. (2020). Manipulation of fruit dry matter via seasonal pruning and its relationship to d'anjou pear yield and fruit quality. *Agronomy*, *10*(6), 897.

Goldberg, N. (2002). *Women are not small men: Life-saving strategies for preventing and healing heart disease in women*. Ballantine Books.

Goodhart, C. A. (1984). Problems of monetary management: the uk experience. In *Monetary Theory and Practice*, (pp. 91–121). Springer.

Graaf, S. v. d. (2018). In waze we trust: algorithmic governance of the public sphere. *Media and Communication*, *6*(4), 153–162.

Grabowski, D. C., Campbell, C. M., & Morrisey, M. A. (2004). Elderly licensure laws and motor vehicle fatalities. *Jama*, *291*(23), 2840–2846.

Grabowski, D. C., & Morrisey, M. A. (2004). Gasoline prices and motor vehicle fatalities. *Journal of Policy Analysis and Management*, *23*(3), 575–593.

Graham, J. D. (1984). Technology, behavior, and safety: An empirical study of automobile occupant-protection regulation. *Policy Sciences*, *17*(2), 141–151.

Greenstein, S. (1993). Markets, standards, and the information infrastructure. *IEEE Micro*, *13*(6), 36–51.

Gupta, R., Tanwar, S., Tyagi, S., & Kumar, N. (2019). Tactile internet and its applications in 5g era: A comprehensive review. *International Journal of Communication Systems*, *32*(14), e3981.

Hasan, S., & Kumar, A. (2019). Digitization and divergence: Online school ratings and segregation in america. *SSRN Working Paper*.

Helfat, C. E. (1994). Evolutionary trajectories in petroleum firm r&d. *Management Science*, *40*(12), 1720–1747.

Hemenway, D. (1975). *Industrywide voluntary product standards*. Ballinger Pub. Co.

Henderson, R., & Clark, K. (1990). Architectural innovation: The reconfiguration of existing product technologies and the failure of established firms. *Administrative Science Quarterly*, (pp. 9–30).

Hertzberg, H. (1970). Misconceptions regarding the design and use of anthropomorphic dummies. Tech. rep., Air Force Aerospace Medical Research Lab.

Hill, K., Edwards, M., & Szakaly, S. (2007). How automakers plan their products: a primer for policymakers on automotive industry business planning. In *Center for Automotive Research*.

Holmstrom, B. (1979). Moral hazard and observability. *The Bell Journal of Economics*, (pp. 74–91).

Holmstrom, B., & Milgrom, P. (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics and Organizations*, *7*, 24.

Hubbard, T. N. (2000). The demand for monitoring technologies: The case of trucking. *The Quarterly Journal of Economics*, *115*(2), 533–560.

Iansiti, M. (1995). Technology development and integration: An empirical study of the interaction between applied science and product development. *IEEE Transactions on Engineering Management*, *42*(3), 259–269.

Ibanez, M. R., & Toffel, M. W. (2020). How scheduling can bias quality assessment: Evidence from food-safety inspections. *Management Science*, *66*(6), 2396–2416.

Ichniowski, C., & Shaw, K. (2003). Beyond incentive pay: Insiders' estimates of the value of complementary human resource management practices. *Journal of Economic Perspectives*, *17*(1), 155–180.

Janssen, E., & Wismans, J. (1987). Evaluation of vehicle-cyclist impacts through dummy and human cadaver tests. In *The 11th Technical Conference on Experimental Safety Vehicles, 12th-15th May*.

Kaczmarska, E., Gawroński, J., Jabłońska-Ryś, E., Zalewska-Korona, M., Radzki, W., & Sławińska, A. (2016). Hybrid performance and heterosis in strawberry (fragaria× ananassa duchesne), regarding acidity, soluble solids and dry matter content in fruits. *Plant Breeding*, *135*(2), 232–238.

Katila, R., & Ahuja, G. (2002). Something old, something new: A longitudinal study of search behavior and new product introduction. *Academy of management journal*, *45*(6), 1183–1194.

Katz, M. L., & Shapiro, C. (1986). Technology adoption in the presence of network externalities. *Journal of Political Economy*, *94*(4), 822–841.

Katz, M. L., & Shapiro, C. (1994). Systems competition and network effects. *Journal of economic perspectives*, *8*(2), 93–115.

Kellogg, K. C., Valentine, M. A., & Christin, A. (2020). Algorithms at work: The new contested terrain of control. *Academy of Management Annals*, *14*(1), 366–410.

Kelvin, W. T. B. (1889). *Popular Lectures and Addresses: Constitution of matter. 1889*, vol. 1. Macmillan.

Kerr (1975). On the folly of rewarding a, while hoping for b. *Academy of Management Journal*, *18*(4), 769–783.

Kerr, Nanda, & Rhodes-Kropf (2014). Entrepreneurship as experimentation. *Journal of Economic Perspectives*, *28*(3), 25–48.

Knight, F. H. (1921). *Risk, uncertainty and profit*, vol. 31. Houghton Mifflin.

Kuhn (1962). *The structure of scientific revolutions*, vol. 111. Chicago University of Chicago Press.

Kuhn, T. S. (1961). The function of measurement in modern physical science. *Isis*, *52*(2), 161–193.

Latour, B. (1986). Visualization and cognition. *Knowledge and society*, *6*(6), 1–40.

Leonard-Barton, D. (1992). Core capabilities and core rigidities: A paradox in managing new product development. *Strategic management journal*, *13*(S1), 111–125.

Lerner, J., & Tirole, J. (2014). A better route to tech standards. *Science*, *343*(6174), 972–973.

Levinthal, & March (1981). A model of adaptive organizational search. *Journal of Economic Behavior & Organization*, *2*(4), 307–333.

Levitt, & Porter (2001a). How dangerous are drinking drivers? *Journal of Political Economy*, *109*(6), 1198–1237.

Levitt, & Porter (2001b). Sample selection in the estimation of air bag and seat belt effectiveness. *The Review of Economics and Statistics*, *83*, 603–615.

Lin, J. Y. (1995). The needham puzzle: why the industrial revolution did not originate in china. *Economic Development and Cultural Change*, *43*(2), 269–292.

Linder, A., & Svensson, M. Y. (2019). Road safety: the average male as a norm in vehicle occupant crash safety assessment. *Interdisciplinary Science Reviews*, *44*(2), 140–153.

Liu, & Mager, D. (2016). Women's involvement in clinical trials: historical perspective and future implications. *Pharmacy Practice (Granada)*, *14*(1).

Liu, Ranjan, B., & Shiller, B. R. (2020). Are coarse ratings fine? applications to crashworthiness ratings. *Working Paper*.

Luca, M. (2016). Reviews, reputation, and revenue: The case of yelp. com. *Harvard Business School Working Paper*.

Luque, B., & Ballesteros, F. J. (2019). To the sun and beyond. *Nature Physics*, *15*(12), 1302–1302.

Maeda, K., & Ahn, D.-H. (2021). Estimation of dry matter production and yield prediction in greenhouse cucumber without destructive measurements. *Agriculture*, *11*(12), 1186.

Manso, G. (2011). Motivating innovation. *The Journal of Finance*, *66*(5), 1823–1860.

March (1991). Exploration and exploitation in organizational learning. *Organization Science*, *2*(1), 71–87.

March, & Simon (1958). *Organizations*. John Wiley & Sons.

Mau, S. (2019). *The metric society: On the quantification of the social*. John Wiley & Sons.

Mazmanian, M., & Beckman, C. M. (2018). "making" your numbers: Engendering organizational control through a ritual of quantification. *Organization Science*, *29*(3), 357–379.

Menon, A. R., & Yao, D. A. (2017). *Rationalizing outcomes: Mental-model-guided learning in competitive markets*. Harvard Business School Working Paper.

Mokyr, J. (2002). *The Gifts of Athena*. Princeton University Press.

Mokyr, J. (2005). The intellectual origins of modern economic growth. *The Journal of Economic History*, *65*(2), 285–351.

Muller, J., & Muller, J. Z. (2018). *The Tyranny of Metrics*. Princeton University Press.

Nelson, R. R. (1985). *An Evolutionary Theory of Economic Change*. Harvard University Press.

NHTSA (1995). National highway traffic safety administration performance report, fy 1994. *Traffic Safety*.

Oberholzer-Gee, F., Yao, D., & Raabe, E. (2006). Goodyear and the threat of government tire grading. *Harvard Business School Teaching Case*.

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

Perez, C. C. (2019). *Invisible women: Exposing data bias in a world designed for men*. Random House.

Porter, T. (1995). *Trust in Numbers*. Princeton University Press.

Prasad, V., Kim, C., Burotto, M., & Vandross, A. (2015). The strength of association between surrogate end points and survival in oncology: a systematic review of trial-level meta-analyses. *JAMA Internal Medicine*, *175*(8), 1389–1398.

Ranganathan, A., & Benson, A. (2020). A numbers game: Quantification of work, auto-gamification, and worker productivity. *American Sociological Review*, *85*(4), 573–609.

Reisch, L., Graulich, K., Degallaix, L., Maurer, S., & Bernefeld, N. (2010). Work on preparatory studies for ecodesign requirements for eups (iii) and on stakeholder representation: Lot c: Stakeholder representation: Consumers-final report. Tech. rep., European Commission.

Rindova, V., Ferrier, W. J., & Wiltbank, R. (2010). Value from gestalt: how sequences of competitive actions create advantage for firms in nascent markets. *Strategic Management Journal*, *31*(13), 1474–1497.

Rosenberg, N. (1969). The direction of technological change: inducement mechanisms and focusing devices. *Economic Development and Cultural Change*, *18*(1), 1–24.

Rosenberg, N. (1982). *Inside the black box: technology and economics*. Cambridge University press.

Rosenberg, N., & Nathan, R. (1994). *Exploring the black box: Technology, economics, and history*. Cambridge University Press.

Rosenkopf, L., & Nerkar, A. (2001). Beyond local search: boundary-spanning, exploration, and impact in the optical disk industry. *Strategic Management Journal*, *22*(4), 287–306.

Roth (2009). Looking at shirley, the ultimate norm: Colour balance, image technologies, and cognitive equity. *Canadian Journal of Communication*, *34*(1).

Roth (2020). Pre-test with caution: Event-study estimates after testing for parallel trends. *Department of Economics, Harvard University, Unpublished manuscript*.

Samaha, R. R., & Elliott, D. S. (2003). Nhtsa side impact research: motivation for upgraded test procedures. In *Eighteenth International Technical Conference on the Enhanced Safety of Vehicles, Paper*, 492.

Sauder, & Espeland (2009). The discipline of rankings: Tight coupling and organizational change. *American Sociological Review*, *74*(1), 63–82.

Scherer, F. (1965). Invention and innovation in the watt-boulton steam-engine venture. *Technology and Culture*, *6*(2), 165–187.

Schneider, L. (1983). Development of anthropometrically based design specifications for an advanced adult anthropomorphic dummy family, vol. 1 final report. Tech. rep., University of Michigan, Ann Arbor, Transportation Research Institute.

Scoville, W. L. (1912). Note on capsicums. *Journal of the American Pharmaceutical Association*, *1*(5), 453–454.

Searle, J., & Haslegrave, C. (1969). Anthropometric dummies for crash research. *MIRA Bulletin*, *5*, 25–30.

Sharkey, A. J., & Bromley, P. (2015). Can ratings have indirect effects? evidence from the organizational response to peers' environmental ratings. *American Sociological Review*, *80*(1), 63–91.

Simcoe, T. (2012). Standard setting committees: Consensus governance for shared technology platforms. *American Economic Review*, *102*(1), 305–36.

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, *69*(1), 99–118.

Simon, H. A. (1973). The structure of ill structured problems. *Artificial intelligence*, *4*(3-4), 181–201.

Spence, M. (1978). Job market signaling. In *Uncertainty in economics*, (pp. 281–306). Elsevier.

Starkey, J., Young, J., Horn, W., Sobkow Sr, W., Sobokow Sr, W., Alderson, S., Cichowski, W., Krag, M., & Auerbach, J. H. (1969). The first standard automotive crash dummy. *SAE Transactions*, (pp. 935–948).

Steen, E. V. d. (2017). A formal theory of strategy. *Management Science*, *63*(8), 2616–2636.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 677–680.

Strathern, M. (1997). "improving ratings": audit in the british university system. *European Review*, *5*(3), 305–321.

Strathern, M. (2000). The tyranny of transparency. *British educational research journal*, *26*(3), 309–321.

Stuart, T. E., & Podolny, J. M. (1996). Local search and the evolution of technological capabilities. *Strategic Management Journal*, *17*(S1), 21–38.

Taylor, F. W. (1911). *Scientific Management*. Routledge.

Timmermans, S., & Epstein, S. (2010). A world of standards but not a standard world: Toward a sociology of standards and standardization. *Annual review of Sociology*, *36*, 69–89.

Utterback (1994). Mastering the dynamics of innovation: How companies can seize opportunities in the face of technological change. *University of Illinois at Urbana-Champaign's Academy for Entrepreneurial Leadership Historical Research Reference in Entrepreneurship*.

Utterback, & Abernathy, W. J. (1975). A dynamic model of process and product innovation. *Omega*, *3*(6), 639–656.

Vasserman, S., Feldman, M., & Hassidim, A. (2015). Implementing the wisdom of waze. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Viegand Maagoe, V. H. e. K. (2018). Review study on vacuum cleaners. Tech. rep., European Commission.

Vinokurova, N., & Kapoor, R. (2020). Converting inventions into innovations in large firms: How inventors at xerox navigated the innovation process to commercialize their ideas. *Strategic Management Journal*, *41*(13), 2372–2399.

Wu, J. (2021). Nothing happens in a vacuum: Can a dominant metric be shifted? *Working Paper*.

Wu, J. (2022). Innovation for dummies? exploring the role of metrics in automotive safety. *Working Paper*.

Yao, D. A. (1988). Strategic responses to automobile emissions control: a game-theoretic analysis. *Journal of Environmental Economics and Management*, *15*(4), 419–438.

Yates, J., & Murphy, C. N. (2019). *Engineering rules: global standard setting since 1880*. JHU Press.

Zelizer, V. A. (1989). The social meaning of money:" special monies". *American Journal of Sociology*, *95*(2), 342–377.