# Quality Disclosure and Regulation: Scoring Design in Medicare Advantage *

Benjamin Vatter†

October 24, 2021

**Job Market Paper**

**Click here for the latest version** and **here for the appendix**

**Abstract**

Regulators often generate quality scores to help consumers with limited information about product quality, as in schooling, healthcare, and financial markets. When designing scores, regulators must not only anticipate how they will influence consumer choices but also the resulting impact on firms' incentives to invest in quality. In this work, I draw on theoretical insights, econometric strategies, and computational methods to develop an empirical scoring design methodology. I apply it to a large health insurance market and find an alternative policy which vastly improves the market's performance. The new design coarsens consumers' information about lower-quality insurance options but refines it for higher-quality ones. Changes to available product information generate a shift in demand towards higher-quality plans, triggering additional firm investments and making consumers better informed about a menu of superior options. The new design also optimally aggregates different quality dimensions, tackling a multitasking moral hazard problem. The friction is due to firms' (agent) private incentives to attain scores using cost-efficient investments instead of consumer-valued ones, preferred by the regulator (principal). Overall, the alternative policy increases welfare by $669 per enrollee per year. The analysis reveals that simple scores can be remarkably effective if well-designed and provides a method to construct them.

---

*I would like to thank David Dranove, Igal Hendel, Gaston Illanes, and Amanda Starc for their invaluable mentorship and advice. I thank Vivek Bhattacharya, Mar Reguant, Robert Porter, William Rogerson, Molly Schnell, Sebastian Fleitas, Jose Ignacio Cuesta, Carlos Noton, Victoria Marone, Matthew Leisten, Eilidh Geddes, and seminar participants at Northwestern University and the NHEBA work group, for their valuable comments and suggestions. This work benefited from generous funding by the Robert Eisner Graduate Fellowship. All errors are my own.

†Northwestern University. Email: benjaminvatterj@u.northwestern.edu

# 1 Introduction

Quality scores are ubiquitous. Certifying agencies use them to grade the quality of schools, the clinical outcomes of hospitals, and the energy efficiency of appliances, among others. Scores help consumers choose when information is scarce and, by doing so, alter firms' incentives to invest in quality. Through these effects, scores can meaningfully impact the market's performance (Dranove and Jin, 2010). Social planners, who play a prominent role as certifiers of quality and designers of scores, are interested in leveraging supply and demand responses to scoring for the market's betterment. However, the endogenous supply responses make this extensive policy problem notoriously challenging, and to date, there is limited empirical evidence on how to optimally design scores.[1]

This paper investigates scoring design empirically by studying the Medicare Advantage (MA) health insurance market. Scores in MA summarize many quality dimensions, including disease management protocols and the quality of providers in each plan's network. Scores are shown to consumers as "stars" ranging from one to five, with half-star increments. The function mapping quality measurements to stars is called the design, which in MA varies yearly. This variation, combined with data availability and extensive quality heterogeneity, makes MA well-suited for the study of scoring design. As plan quality impacts mortality (Abaluck et al., 2021) and public spending (CMS, 2016), it is also a setting where improvements might have significant welfare consequences. I evaluate the potential impact of redesigned scores by developing and estimating a model of the market and solving the designer's problem using a novel methodology. The results indicate insurers underinvest in quality relative to an efficient full-information social optimum, and substantial surplus is lost due to consumers' limited information. Moreover, there is pervasive misclassification wherein consumers would prefer, ex-post, some lower-scoring products to higher-scoring ones. I develop a new design, revealing mechanisms through which scores can address these distortions. The exercise shows that scoring design might significantly improve welfare in various markets with an endogenous provision of quality and limited consumer information.

Designing welfare-improving quality scores requires overcoming a series of hurdles. First, it is unclear how much quality information to share with consumers. If insurers respond to scores by adjusting quality, then it is no longer the case that a more informed consumer is better off. For example, a system awarding binary certifications indicating whether a plan is high-quality or not reveals less information than one with an additional level for medium-quality. However, replacing

[1]The main difference between this work and the theory on information design with endogenous quality is that my setting includes competition among firms both in quality and prices. Quality is also multidimensional, as it is in many empirical settings. The closest theoretical paper to this work is Zapechelnyuk (2020), which studies surplus-maximizing scoring design for a scalar quality provided by a monopolist. The theory is further discussed later in the introduction and throughout the paper.

the first design with the second might lead some insurers to reduce their investments and decrease welfare.[2] Second, evaluating the optimality of any score requires knowing the social value of quality, its production cost, and how changes to information affect prices and competition. Finally, using these inputs to design new scores involves solving a policy design problem for which there are no known systematic solution approaches or optimality conditions to leverage.[3]

To make progress on these challenges, I begin by documenting how variation in the design of the MA scores has affected demand and quality in the market. Using variation produced by the introduction of new quality dimensions to the MA scoring system (e.g., breast cancer screening rates), I provide novel evidence that insurers adjust their quality in response to scoring incentives. The analysis leverages how the regulator transforms continuous quality into discrete scores, creating heterogeneous incentives for improvements across firms. In particular, the system does not award additional stars for improvements made by high-quality plans but penalizes low-quality plans that fail to improve. I use this feature to create a triple-differences strategy, showing insurers respond to design changes by increasing their quality substantially, and more so if they risk low scores. Therefore, quality in MA reacts to scoring incentives as it does in other markets (Jin and Leslie, 2003; Barahona et al., 2020), and the designer faces an endogenous supply response and its associated challenges.

The reduced form results also show consumers prefer higher scoring products and respond to scores differently depending on how these are designed. For example, all else equal, consumers prefer plans with four stars to those with three. They value the difference between scores more when the system demands greater improvements in medical outcomes to achieve the increase, as opposed to improvements in chronic condition management. The evidence contributes to extensive literature documenting demand responses to the MA scores (Dafny and Dranove, 2008; Reid et al., 2013; Darden and McCarthy, 2015). Overall, the supply and demand responses to scoring variation contain information about the social value of quality and its production cost, which are fundamental inputs to the scoring design problem. To disentangle this information from the overall data variation, I develop and estimate a model of the market.

In the model, a regulator first announces a public scoring rule. Insurers then invest in multiple quality dimensions, such as contracting with better hospitals to reduce readmissions rates or with more pharmacies to increase vaccinations. Following, insurers compete in prices over a collection of products, and consumers choose among them. Their choice maximizes expected utility subject

[2]For example, in markets with a unique, efficient level of quality production, dichotomous certification can lead to efficient investments (see Section 2). A three-level score, instead, might lead firms to differentiate in quality to reduce price competition, thus decreasing welfare, as in Ronnen (1991).

[3]The technical complexity of scoring design is that it is a functional optimization problem over a space of discontinuous functions that map multiple dimensions down to a few scalars.

to uncertainty over quality, partially informed by each product's score. This model builds on those of Curto et al. (2021a) and Miller et al. (2019), extending them to a setting of multidimensional quality competition with an incompletely informed demand. Moreover, I prove that scoring design variation and plan enrollment data identify consumers' preferences and beliefs over quality. The rationale is that the reduced-form finding that consumers are willing to pay more for the same increase in scores under different designs reveals how much they value each quality dimension. I estimate the model using individual-level data on enrollment choices, product quality, and the MA scoring rules.

The estimates suggest quality and information are inefficiently supplied in MA. Holding prices and qualities constant, consumers lose $185.9 in surplus per year because scores pool heterogeneous products and often classify plans consumers prefer less as having higher quality. Misclassification occurs because scores aggregate quality dimensions differently from consumers' preferences: consumers might prefer a plan that excels in medical outcomes over one which does so in diagnostic services, but the system might assign a higher score to the second. The estimates also show consumers would prefer if plan quality was on average higher and would be willing to pay more than what it would cost to produce. Insurers, however, have no incentive to invest because improvements would be hidden from consumers due to the coarseness of the scores.

I use the estimated model and a novel methodology to find welfare-maximizing scores. I restrict attention to a class of designs that deterministically assign higher scores to greater quality and use finitely many scores. This class incorporates many standard disclosure policies, such as the MA scores, car-safety certifications, and school quality grades.[4] I show these designs are a composition of a continuous aggregator that combines quality dimensions into an index and a cutoff function that segments the index into scores. I derive a novel solution approach to the scoring design problem by combining this decomposition result with the insight of Kamenica and Gentzkow (2011) that choosing a disclosure policy is akin to selecting a distribution over consumers' posterior beliefs. The approach consists of first computing a large set of counterfactual equilibria and then identifying the value of every alternative design as a distribution over these.[5] Using the new approach, I find a (constrained) optimal scoring design for MA that uses the same data and number of scores as the current system, making it a direct substitute.

The new design reveals total welfare could be improved by $669 per Medicare beneficiary per year, or about a month's worth of subsidies for the entire MA population. The alternative

[4] Dworczak and Martini (2019) prove that similarly defined scores can be optimal in a wide array of scenarios, albeit with exogenous quality. Their definition allows for specific segments of the score to be fully revealing. I explore these designs in the appendix.

[5] This approach shares some similarities with other grid methods from the Industrial Organization literature, such as importance sampling (Ackerberg, 2009) or discretizations of continuous distributions (Fox et al., 2011).

system pools a wide array of low qualities into a single scores and sets a high standard for getting the ones signaling high quality. Consumers penalize products getting low scores, buying better scoring ones instead. The shift in demand creates incentives for firms to invest enough to avoid receiving the low-quality score, thus offsetting the underprovision of quality in MA and improving welfare.[6] The new design also changes how quality dimensions are aggregated, improving the alignment with consumers' preferences, thus reducing misclassification. The change also tackles a related multitasking moral hazard problem (Holmstrom and Milgrom, 1991), pervasive to scores that summarize multiple dimensions of quality, such as in nursing homes (Feng Lu, 2012), energy-efficiency (Clay et al., 2021), and schools (Neal and Schanzenbach, 2010). Intuitively, every scoring system creates different investment paths for plans to reach a score, but firms and the regulator might disagree over which to take. For example, a plan could attain four stars by having good hospitals in its network and providing excellent diagnostic services or having the best hospitals and average diagnostics. Insurers might find the first alternative cheaper, while consumers and the regulator might prefer the second path. Changing the scores' aggregation method controls firms' options and addresses this problem. Overall, about half of the welfare gains from the new proposed design come from improving quality investments, both in their total amount and relative allocation across dimensions.

The second half of the welfare gains stem from consumers being better informed about quality. The analysis shows that adding or subtracting scores from the system introduces a trade-off. Additional scores allow heterogeneous products in the market, which is valuable to consumers unwilling to pay the price tag of a high-end product, and for firms which cannot efficiently produce high-quality. The downside is that adding additional scores reduces firms' incentives to invest efficiently. Intuitively, coarse scores penalize firms deviating from efficient production by pooling them with worse products. Added scores weaken the penalty by making the worse product in the same pool better. Therefore, how many signals, or distinct stars, are optimal for a given market depends on how heterogeneous consumers are in their willingness to pay for quality, how varied investment costs are, and the welfare cost of inefficient production.

The analysis of scoring granularity also reveals that simple scores can be remarkably effective. I compare the case of binary quality certifications – which are the simplest and most common type of quality scores – against a fully informative scoring system.[7] While a fully informative score would

[6] Harbaugh and Rasmusen (2018) show that coarse information is optimal due to similar logic in a theoretical framework with exogenous quality and voluntary participation.

[7] Common examples of certification are the USDA or EU organic labels, front-of-package warning labels for sugars and calories, Energy Star certification on computers and monitors, NAHQ certification of medical professional quality, and a plethora of ISO certifications ranging from upper management to food processing technology quality. Fully informative designs are infeasible when aggregating multiple continuous dimensions into a single score unless consumers have common preferences over quality.

undoubtedly help consumers choose, it might exacerbate other frictions such as market power over quality (Crawford et al., 2019). In particular, a firm that can freely choose and charge for its quality to informed consumers might invest differently than socially optimal. The reason is profits do not fully reflect the social value of investments (Spence, 1975; Ronnen, 1991), which in MA I find would lead to underproduction – insurers underprovide quality in the status quo and would continue to do so even under full information. In contrast, a certification that leads enough consumers to avoid uncertified products effectively restricts firms' choices to either meet the certification standard or not invest at all. Anything in between is wasteful as firms cannot charge for their investments.[8] As a consequence, I find that a well-designed certificate can outperform a fully informative score both in quality output and welfare. Thus, even if consumers were capable of processing complex quality information, regulators might still prefer simple scores.

Finally, I study how variations to the regulator's objective and information affect the design. First, the regulator might have private preferences for quality because they are better informed about the health consequences of different insurance choices. However, I show that using scores to nudge consumers towards specific plans rapidly erodes the scores' informational value, firms' incentives to invest, and welfare. Thus, quality scores are a poor nudging mechanism. Second, I examine scoring design with limited information about consumers' quality preferences.[9] I derive an alternative, robust design approach that maximizes welfare under the worst-case preferences and can be solved using my methodology with an added linear constraint. The resulting design improves welfare and operates through the exact mechanisms as the main results: a high standard for top scores jolts investments, and quality aggregation aligns firms incentives with the regulator's preferences. Notably, the existing aggregation scheme is effective in this scenario, suggesting that the MA scores' designer might be uncertain about consumers' quality preferences and averse to misrepresenting them. My work proves formal conditions under which these preferences are identified and delivers a method for their estimation.

Overall, I contribute to a growing literature studying the design of disclosure policies. By solving a constrained optimal design problem empirically, I bridge the gap between the theoretical literature searching for optimal designs (Albano and Lizzeri, 2001; Rodina and Farragut, 2016; Boleslavsky and Kim, 2018; Ball, 2019; Zapechelnyuk, 2020) and the empirical literature measuring scores' impact (Bollinger et al., 2011; Werner et al., 2012; Elfenbein et al., 2015; Chen, 2018; Araya et al., 2018; Houde, 2018b).[10] My analysis incorporates firms' responses to counterfactual scores,

[8] Zapechelnyuk (2020) shows that the connection between scores and delegation problem extends beyond simple certifications, and in some cases, can be used to solve the scoring design problem.

[9] The regulator would have uncertainty about quality preferences if consumers did not understand the existing design variation. In this case, I show, consumers' preferences might only be set-identified.

[10] See Dranove and Jin (2010) for a review of earlier work on quality disclosure. The related theory literature is also linked to the work in information design and Bayesian persuasion, reviewed in Kamenica (2019).

which contributes to the research on the supply effects of centralized mandatory disclosure, studied in education (Mizala and Urquiola, 2013; Allende et al., 2019), health care (Chou et al., 2014), airlines (Forbes et al., 2015), and electrical appliances (Houde, 2018a).

The regulatory role of scores connects this study to the literature on quality provision and regulation. The origin of inefficient quality in imperfect competition has been studied theoretically (Spence, 1975; Mussa and Rosen, 1978; Schmalensee, 1979; Ronnen, 1991), and documented empirically (McManus, 2007; Crawford et al., 2019). I contribute to this evidence and the literature on the provision of quality in healthcare markets (Cutler et al., 2010; Cooper et al., 2011; Gaynor et al., 2013; Kolstad, 2013) by showing that insurers can alter the quality of their services and choose to underprovide it relative to the social optimum. My results indicate indirect quality regulation using scores can offset this market failure which contributes to the empirical study of quality regulation broadly (Angrist and Guryan, 2008; Larsen, 2014; Larsen et al., 2020; Barrios, 2017; Kleiner and Soltas, 2019; Farronato et al., 2020; Atal et al., 2021). To study how firms respond to changes in regulation, I build on research studying competition over non-price product attributes (Berry and Waldfogel, 2001; Gandhi et al., 2008; Nosko, 2014; Fan, 2013; Berry et al., 2016; Hui et al., 2018; Fan and Yang, 2020) and among health insurers (Dafny, 2010; Ho and Lee, 2017; Ryan, 2020; Ho and Handel, 2021). I contribute to these strands by studying multidimensional quality competition among heterogeneous insurers with incompletely informed demand.

Finally, I provide a policy contribution in the form of a new scoring design, implementable with the same inputs and technology currently in use, and which would increase welfare by over $43 billion per year. The contribution builds on a broad literature studying the industrial organization of MA (Town and Liu, 2003; Lustig, 2010; Aizawa and Kim, 2018; Curto et al., 2021a; Nosal, 2011; So, 2019; Charbi, 2020; Miller et al., 2019). To my knowledge, this is the first paper to study the scoring design policy question empirically.

I organize the remainder of the paper as follows. Section 2 illustrates the regulatory role of disclosure policy using the single-product monopolist model of Spence (1975). Section 3 presents the setting and describes the MA Star Rating system. Section 4 studies the effects of these scores on demand for insurance and the supply of quality. The findings clarify the potential for MA scores to regulate quality and the variation leveraged in the structural model. Section 5 presents the structural model. Section 6 describes how I estimate the model and how the data variation identifies it. Section 7 develops the main analysis showing the alternative scoring design and how variations affect welfare in the market. Section 8 discusses the robustness of the findings to alternative assumptions about the designer's information and constraints. Section 9 concludes.

## 2 Quality Disclosure and Regulation

I begin by describing the economic intuition underlying scores' ability to inform consumers and regulate quality simultaneously.[11] Following Spence (1975), consider a monopolist that chooses a quality $q$ and price $P$ for its single indivisible product. Consumers decide whether to buy a unit of the good, resulting in an inverse demand $P(x, q)$, where $x$ denotes quantity. The monopolist's production cost is $C(x, q)$ and its profits are $\pi(x, q) = xP(x, q) - C(x, q)$. The market's total welfare is

$$W(x, q) = \int_0^x P(v, q)dv - C(x, q)$$

For any quantity, the monopolist chooses a quality that equates its marginal revenue to its marginal cost ($xP_q(x, q) = C_q(x, q)$). Welfare, instead, is maximized when the marginal cost of quality equates its marginal surplus ($\int_0^x P_q(v, q)dv = C_q(x, q)$). The quality that satisfies one condition, in general, fails the other. In part, this is because the monopolist does not value the impact of its investments on infra-marginal consumers. As Spence notes, depending on how the marginal and average consumers compare in their willingness to pay for quality, the monopolist will over or underprovide quality relative to the social optimum.

The solid lines in Figure 1a present the profit and welfare curves for one such monopolistic market. In this scenario the monopolist would underprovide quality despite market pressures and is, therefore, said to have market power over quality (Crawford et al., 2019). This power results in a wedge between the efficient ($q^W$) and the profit-maximizing choices ($q^*$), known as the *Spencian* distortion.

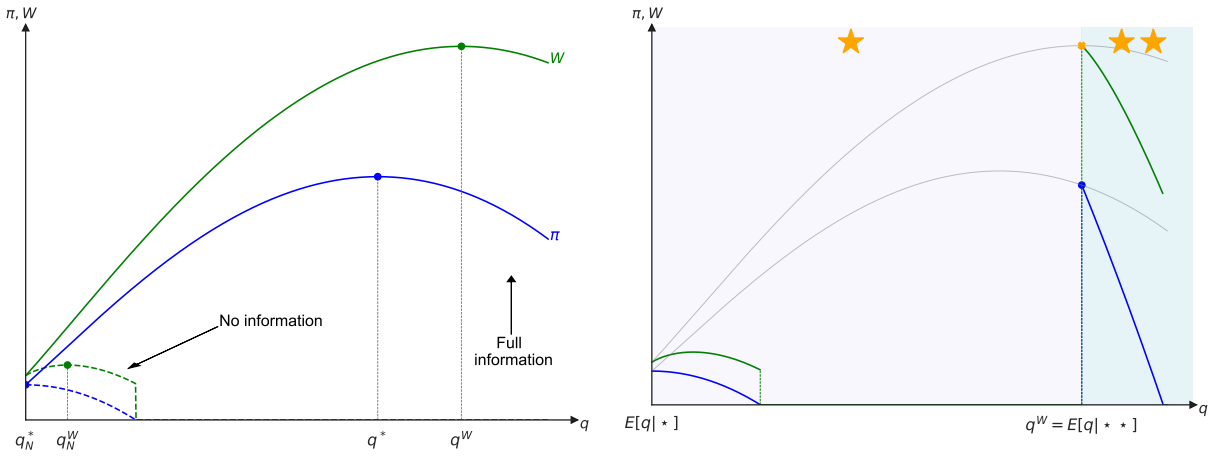
Now suppose consumers cannot ascertain the product's quality before purchase but have rational expectations.[12] Lacking any information, consumers rationally expect the monopolist to provide his cost-minimizing quality. The monopolist, in turn, does precisely that, as it is unable to charge for its investments to uninformed consumers. Figure 1a depicts the resulting scenario in dashed curves.[13]

A regulator that observes this failure might choose to intervene by creating a quality scoring system. They would first announce a rule by which they will assign quality measurements to a set of scores. The monopolist would then make its investments and pricing decisions, and the

[11] To focus on the primary forces used by the designer, I abstract away from unobserved investments, multidimensional quality, and other regulatory concerns found in the application.

[12] I assume that there are no alternative mechanisms to achieve the full-information outcome. For example, the monopolist cannot credibly commit to producing a certain quality, or the market is short-lived.

[13] No other belief can be held in a rational expectations equilibrium as the monopolist always has incentives to reduce quality to save on costs. As consumers cannot observe this deviation, the demand for the product remains the same, increasing the monopolist's profits.




(a) Informational extrema (b) Regulated

Figure 1: Quality certification under monopolistic provision

*Notes*: This figure illustrates how quality certification changes a monopolist's incentive to invest in quality. On the left, the figure presents the profit and welfare curves for the full and no information scenarios. The figures under no information vanish at the point in which the monopolist makes negative profits and exits the market. On the rights, the figure illustrates how the profit and welfare curves change when consumers are only informed that quality exceeded the socially optimal level. The shaded areas illustrate the distinct scores.

regulator would measure quality and disclose the appropriate score to consumers. For example, the regulator could recover the full information scenario by making the scores equal to the measured quality, eliminating information asymmetries but restoring the Spencian distortion.

The fundamental insight behind scores' ability to improve welfare beyond full information is that, conditional on quality, scores change the firm's demand but not the product's value to consumers. Scores can, therefore, align the monopolist's incentives with the regulator's objective. For example, Figure 1b illustrates the outcome of a policy showing consumers one star if quality is underprovided and two if it is efficient or overprovided. This policy disrupts the firm's profit curve as to the left of the two-star cutoff ($q^W$), consumers have the same beliefs as under no-information. To the right, consumers also rationally expect the quality to be cost-minimizing but know that it at least exceeds the cutoff. On both extremes, the modified profit and welfare curves begin at their full-information value. The regulated monopolist will choose to provide the efficient quality, consumers' will believe it to be as such, and the regulator eliminated the spencian distortion at no informational cost to consumers.

This illustration reveals that the regulatory power of scores stems from their ability to marshal demand. By disclosing a coarse signal, the regulator coordinates consumers to demand high quality, offsetting the monopolist's market power. While, in general, this effect comes at a loss of information for consumers, the intuition that this coordination can lead coarse scores to outperform full information extends well beyond this simple example. In Appendix 2 I show that it applies

to monopolistic markets in general, including if the monopolist overprovides quality. I also show that the intuition applies if the regulator has limited information about the monopolist's costs, as in Zapechelnyuk (2020), or if consumers update a Bayesian prior instead of having rational expectation. The solution often differs from a binary system, as the regulator faces a trade-off between informing consumers and controlling quality. Detailed scores give consumers more information and accommodate product heterogeneity at the expense of decreasing firms' incentives to invest. To determine the optimal design for a given market, we must uncover firms' investment costs and the social value of quality. These will determine the socially optimal investment for each firm, which we can then attempt to generate with a coarse scoring system. To do so, I develop an empirical methodology to recover these components and a method to systematically translate these inputs into optimal scoring designs.

# 3 Institutional Details and Data

## 3.1 Medicare Advantage and The Star Rating Program

Since 1965, retirees and disabled individuals in the US have had access to a public health insurance system known as Medicare. This system provides hospital, physician, and outpatient coverage under a publicly administrated and highly subsidized scheme. A series of reforms enacted between 1982 and 2003 established an alternative to traditional Medicare (TM), known today as Medicare Advantage (MA). Under MA, the Center For Medicare and Medicaid Services (CMS) contracts with private insurance companies to provide alternative coverage for Medicare beneficiaries in exchange for a prospective risk-adjusted capitated payment. Enrollment in MA has been steadily increasing during the past decade; out of the nearly 65 million Medicare-eligible enrollees in 2019, 34% chose a plan in MA.[14]

MA markets are highly concentrated and regulated. During 2019, the average market (county) had 90% of its enrollment controlled by only two firms. At the national level, four firms command 69% of all enrollment (Frank and McGuire, 2019). In most counties, insurers offer a wide array of plans differing in their coverage generosity (e.g., coinsurance, deductibles) and their access to clinical quality. CMS strictly regulates the financial characteristics of plans, including minimum requirements on coverage generosity and limits on premiums relative to coverage. Curto et al. (2021b) provide further description of price and coverage regulation, which I complement with extensive detail in Appendix Section 3.1. CMS also subsidizes consumers by paying a large fraction

[14]Traditional Medicare is composed of part A (hospital coverage) and part B (physician and outpatient coverage). For further details on the history of this program, see McGuire et al. (2011). For details on risk-adjustment and residual selection, see Brown et al. (2014) and So (2019).

of plan premiums. Nearly half of all MA plans are offered at a zero premium to consumers.[15]

In contrast to financial characteristics, differences in plan quality are less regulated and harder for consumers to ascertain. These differences are due to variation across plans in the size and makeup of provider networks, disease management protocols, and processes for approving costly medical procedures, among other factors. Information regarding these aspects of plans are rarely available to consumers when choosing coverage, and when they are, it is often in the form of technical documents that require specialized knowledge to parse. Therefore, to assist consumers, CMS created the MA Star Ratings.

Displayed next to the enrollment button in Medicare's unified shopping platform, the Star Ratings provide a coarse summary of the quality of each plan.[16] To compute these scores, CMS first collects information on over sixty measures of quality for each plan and categorizes them into five groups: Outcome (e.g., readmission rate), Intermediate Outcomes (e.g., diabetes management), Access to Care (e.g., management of appeals), Patient Experience (e.g., customer service), and Process (e.g., breast cancer screenings). Having collected the data, CMS assigns a discrete measure-level score of one to five to each plan-measure, ascending in quality. Next, CMS computes a continuous score for each plan by choosing a weight for each category and computing a weighted average of all measure-level scores for each plan. Overall, denoting $\mathcal{K}$ the set of quality categories, and $w_k$ the weight that each category $k \in \mathcal{K}$, and $\mathcal{L}_k$ the measurements included in the category, the score of plan $j$ is given by

$$\text{Score}_j = \text{Round}_{.5}(\underbrace{\frac{\sum_{k\in\mathcal{K}} w_k \sum_{l\in\mathcal{L}_k} \text{MeasureScore}_l(q_{lj})}{\sum_{k\in\mathcal{K}} w_k |\mathcal{L}_k|} + \omega_j}_{\text{Continuous Score}}) \tag{1}$$

Where $\text{Round}_{.5}(\cdot)$ rounds a number to its nearest half and $q_{kj}$ is the quality of plan $j$ in measure $k$. In the expression $\omega_j$ is a combination of several regulatory features which I call the adjustment factor. I provide further details about the exact design in Appendix 3.1.1.[17]

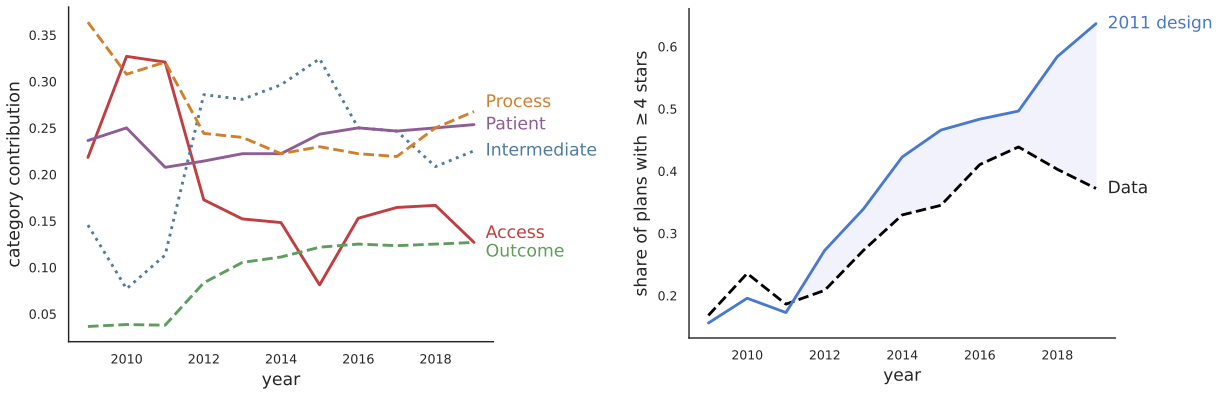
CMS has frequently changed the weights and number of measures in each category, introducing substantial variation in the Star Rating design.[18] In 2012 CMS moved from uniform weights to a

[15] MA consumers still pay their part B premiums. However, this amount is paid regardless of their choice of TM or MA.

[16] See Appendix figure 1 for a view of the platform. The Star Rating program has evolved in its form over the years. For a description of earlier designs, see Dafny and Dranove (2008).

[17] CMS computes the scores at the contract instead of the plan level. Contracts are aggregations of multiple plans of the same insurer that share the same quality. Approximately, contracts define the network and quality of an insurance product while a plan determines the cost-sharing attributes of the product. All the terms entering equation (1) are at the contract level.

[18] CMS also varied the measure-level scoring function over the years, although their yearly change is modest. Ap-




(a) Design variation

(b) Variation in score assignment

Figure 2: Scoring design variation and simulated effect under fixed quality

*Notes*: Figure (a) shows the change in the contribution of each category to the overall score. The contribution of a category corresponds to the product of its number of measurements (e.g., Process includes breast cancer screening and kidney disease monitoring) and its weight, divided by the total weight among all measurements. Figure (b) shows the change in scoring assignment that would result if CMS had kept its 2011 scoring design, keeping quality as measured in the data. I select 2011 as the baseline year because that year's design shares a large number of measures with both previous and following years. The shaded area highlights the gap across the resulting assignments. Adjustment factors are preserved as measured in the corresponding year.

design that gives each Outcome and Intermediate Outcome measure three times the weight of any Process measure and twice the weight of any Access or Patient Experience measure. The size of each category changed each year as CMS introduced and removed measures from each. Overall, each category's contribution to the scores has varied significantly, as shown in Figure 2a. As I detail Appendix 3 , this design variation was likely observed by consumers, as the composition of categories was visible on the Medicare website and in booklets sent by CMS to enrollees.

The design variation significantly impacted score assignment. For example, increasing the share of enrollees receiving adequate care for their high blood pressure (an Intermediate Outcome measurement) from 45% to 65% was 86% more valuable for scoring purposes in 2012 as in 2011. In contrast, increasing breast cancer screening (a Process measure) from 60% to 75% was 38% less valuable.[19] To illustrate the overall change in score assignment, Figure 2b shows that if CMS had kept the 2011 scoring design, 60% instead of 40% of 2019 plans would have received four or more stars. The gap between the two systems is largely due to a decrease in the importance of Access and an increase in that of both Outcome categories. Thus, in 2011 a high-quality plan was one that afforded consumers great access to physicians and a median-quality network of hospitals, while in

pendix 3.1.1 provides further details.

[19] Adequate care here means members with high blood pressure received treatment and were able to maintain a health pressure. Breast cancer screening rates are among women 40 to 69. The scoring value is in terms of the continuous score.

2019 the roles of hospital quality and access to physicians were reversed. The figures also show an improvement in overall quality as the share of top-rated plans increases regardless of the design.

Because insurers can offer the same network arrangement and services under different cost-sharing and premium combinations, CMS measures quality at a level slightly larger than a plan. A contract is a grouping of plans that share the same quality and insurer. The median contract has only two plans, with 70% of its enrollment in one of them. Consumers observe a score for each plan and often will have only one of a contract's plan available in their county. Therefore, in many cases, the distinction between plan and contract is irrelevant. However, having price variation conditional on quality and year will prove useful for estimation purposes. Throughout, I refer to products in MA as plans and refer to contracts only when relevant for clarity or exposition.

Finally, CMS provides dynamic incentives. Starting in 2012, the rebate share and benchmarks of plans vary with their score in the previous year, and the adjustment factor ($\omega_j$) rewards quality improvements.[20] However, this paper aims to understand the short-run mechanisms, effects, and design of a purely informational quality disclosure policy. Thus, in the model and estimations that follow, I incorporate these dynamic features as they appear in the data and treat them as sources of revenue heterogeneity. I do not include pecuniary incentives as part of the designer's toolkit to avoid confusing gains from information design with those from direct transfers.

### 3.2 Data

This paper combines three data sources, the first being plan-market level data from 2009 to 2019. Each year, CMS publishes a compendium of data sets containing information on each MA plan in each county. I use it to construct a panel of plans with their market-level enrollment, benchmarks, prices, rebates, premiums, and the full detail of plan benefits and cost-sharing. Additionally, the data provides the total number of Medicare eligible beneficiaries in each county, and information regarding dual Medicare-Medicaid eligible population. I use these data to adjust the sample, removing dual eligibles and plans specifically designed for that population.[21] I present the descriptive statistics of the data and details regarding their construction in Appendix 3.2.

The second data source is the Medicare Current Beneficiary Survey (MCBS). This nationally representative rotating panel tracks around 15,000 Medicare beneficiaries each year, for up to four year. I obtain the data for 2009 to 2015, which provides me with information on individual demo-

[20] MA also rewards plans achieving five stars by allowing consumers to switch into them after the open enrollment period ends. As there are few five-star plans, I exclude this behavior from the analysis by only considering demand within the open enrollment period. Another dynamic behavior not treated in this article is contract consolidation. This was a practice exploited by few insurers to combine contracts to manipulate their scores for one year. I discuss this further in Appendix 3.1.

[21] Similar data restrictions have been used by Aizawa and Kim (2018), Miller et al. (2019) and Curto et al. (2021a).

graphic, well-being, income, location, and self-assessed knowledge about the Medicare program.[22] Most importantly, it provides plan choices that can be linked with the aggregate data. I restrict the data to non-dual beneficiaries, within the continental US, and with geographic information, leaving 46,833 beneficiary-years. Importantly, the MCBS provides sampling weights to compare the survey's demographics with the national population. However, because of the limited size of these data, they do not include all counties. This will limit my final welfare measurement to about a third of the overall population of Medicare in 2015, or about 22 million individuals.

Finally, the third data source pertains to the quality of plans and the scoring rules. Each year, CMS publishes the data used to compute the star ratings. These files contain CMS's quality measurements and their associated star rating and cutoffs. The challenge in using these data is that the measurement and scale of the underlying quality dimensions have changed over time. Additionally, the files do not contain the direction of improvement and range for dimensions measured but with zero weight ($w_k = 0$). To tackle this challenge, I complement the data by reviewing a decade of CMS communications to insurers regarding scoring design changes. From this, I fully recover the missing information on measures and uncover year-to-year changes to the scoring design. I review these rules in Appendix 3.1.1.

# 4 Demand and Quality Responses to Scoring

The score's effect in the monopoly regulation example of Section 2 relied on two fundamental market behaviors. First, consumers understood the scoring design and adjusted their beliefs accordingly. Second, the monopolist responded to the incentives created by the design and modified its quality. This section explores whether variation in the MA scores produced similar responses among consumers and insurers.

## 4.1 Demand Responses

Consumers might respond to the scoring policy in two ways. For any given year, they might prefer higher- to lower-scoring plans because scores signal quality. Fixing the year means the same design applies to all products. In contrast, as the years pass, the regulator changes category weights in the design, affecting the interpretation of scores. For example, a plan excelling in Medical Outcome quality while being average at everything else would have received three stars in 2011 and four the following year, despite its quality not changing. What changed is the weight on Medical Outcomes and, therefore, the meaning of having four stars. Consumers might respond to the variation in

[22] The MCBS for 2014 is not included as it was never released to the public because of implementation difficulties.

policy by valuing scores differently depending on how they reflect specific category-level quality.

The preference of MA consumers for higher-scoring products is well documented (Dranove and Dafny, 2008; Reid et al., 2013; Darden and McCarthy, 2015). I contribute to the evidence by using my individual-level data to control for two potentially confounding effects. First, plans might be heterogeneous in dimensions beyond their financial characteristics and scored quality. If a plan's quality is positively correlated with consumers' unobserved preference for it, a simple analysis might inflate the effect of scores on demand. I account for this source of bias by studying demand changes within a contract. Second, consumers in MA have substantial switching costs (Nosal, 2011), potentially due to the hassle of changing primary care physicians and interruption in treatments. Thus, some consumers might fail to switch away from a plan whose quality deteriorates, dampening the apparent effect of scores on the plan's market share. I control for this confounding factor by restricting the analysis to consumers not previously in MA.

To study the effect of scoring assignments on demand, I regress consumers' plan choices on scores, product and consumers' characteristics, and contract and market fixed-effects, following the specification:

$$y_{ijt} = \sum_{r=1}^{5} \alpha_r \mathbb{1}\{r_{jt} = r\} + \gamma_{c(j)} + \mu_{m(i)} + \xi_t + \boldsymbol{x}_{ijt}\boldsymbol{\lambda} + \epsilon_{ijt}$$

Above, $y_{ijt}$ indicates whether consumer $i$ chose plan $j$ in year $t$, $\mathbb{1}\{r_{jt} = r\}$ indicates if the plan's score is $r$, and $(\gamma_{c(j)}, \mu_{m(i)}, \xi_j)$ are fixed-effects for the plan's contract, the consumer's market, and the year, respectively.[23] Additional controls, $\boldsymbol{x}_{ijt}$, include demographic variables, such as education, self-assessed health status, ethnicity, gender, disability status, and plan characteristics, such as prescription drug coverage and additional plan benefits, such as dental coverage.

The first column of Table 1 presents the estimates of $\alpha_r$. The results show that improving a plan's score significantly increases its demand. For example, an improvement from four to five stars roughly increases the choice probability of a plan by 0.6% relative to all options, which corresponds to a 31.6% increase in demand relative to a four-stars plan. The gap between the two numbers is caused by new MA enrollees having both a very large set of plans to chose from and a high probability of choosing TM. In this restricted sample, the average MA plan is chosen only 1.58% of the time. Nevertheless, the effect of scoring assignment on demand is significant and meaningful for insurers' revenues.

Having documented consumers' preferences for scores, I now turn attention to how these

[23] As insurers can offer the same insurance contract under different cost-sharing combinations and labels, I do not control for the label-indicator of plan in $\gamma_j$, but rather for the higher-level contract name. Contracts are insurer specific and are the level at which quality is measured by CMS. To reduce the number of effects to estimate in this limited sample, I round half-stars up to their nearest integer.

Table 1: Demand Responses to Scoring

| | I | | II | | III | |
|---|---|---|---|---|---|---|
| **Rounded star rating ($\alpha_r$)** | | | | | | |
| 2 stars | -0.000 | (0.003) | -0.035*** | (0.006) | -0.033*** | (0.008) |
| 3 stars | 0.001 | (0.001) | -0.036*** | (0.004) | -0.027*** | (0.003) |
| 4 stars | 0.007*** | (0.002) | -0.043*** | (0.004) | -0.030*** | (0.003) |
| 5 stars | 0.012*** | (0.003) | -0.036*** | (0.006) | -0.023*** | (0.005) |
| **Rating category weight ($\beta_r$)** | | | | | | |
| 2 stars | | | 0.728*** | (0.116) | 0.205*** | (0.047) |
| 3 stars | | | 0.770*** | (0.077) | 0.189*** | (0.019) |
| 4 stars | | | 0.881*** | (0.083) | 0.219*** | (0.021) |
| 5 stars | | | 0.831*** | (0.094) | 0.204*** | (0.025) |
| **Category** | | | Outcome | | Intermediate | |
| N | 421606 | | 421606 | | 421606 | |
| $R^2$ | 0.712 | | 0.713 | | 0.713 | |

*Notes*: This table displays the estimates of the demand response to rating and scoring design. Only two categories are shown for space, the full table is shown in Appendix Table 4. Observations are weighted by the MCBS sampling weights. The omitted category are new plans and plans that don't have star ratings due to insufficient enrollment in previous years. Standard errors in parentheses are heteroskedasticity robust. *p<0.05, **p<0.01, ***p<0.001.

change with the design. Specifically, I add the contribution of category $k$, shown in Figure 2a, to the previous regression. The contribution ($W_{kt}$) is defined as the total weight of all measurements of category $k$, relative to the total weight of all measurements in all categories.[24]

$$y_{ijt} = \sum_{r=1}^{5} (\alpha_r + \beta_r W_{kt}) \mathbb{1}\{r_{jt} = r\} + \gamma_{c(j)} + \mu_{m(i)} + \xi_t + \boldsymbol{x}_{ijt}\boldsymbol{\lambda} + \epsilon_{ijt}$$

Columns II and III of Table 1 show that the estimated coefficients for $\beta_r$ are statistically significant for all categories. The results indicate that a score improvement was more valuable in later years when the design placed greater weights on the Outcome and Intermediate Outcome categories. In those years, plans with greater scores were more likely to have better quality in these categories. The results indicate that increasing the contribution of Outcome quality to the scores by 1% increases the choice probability gained by a plan improving from three to four stars by 0.1%. This improvement is roughly a 7% increase in the demand for the improving plan.

These design effects also serve to test whether consumers are informed of the scoring rules.

[24]Formally, $W_{kt} = (|\mathcal{L}_{kt}|w_{kt})/(\sum_{k' \in \mathcal{K}} |\mathcal{L}_{k't}|w_{k't})$.

Under the hypothesis that consumers are completely ignorant of the rules, the change in demand when a product increases in scores should be independent of how that score was calculated. Finding that consumers value increases differently when certain categories are more represented in the design rejects this hypothesis. The finding also agrees with the institutional features of MA, as the weights are largely given by the number of measurements included in each category, which are visible to consumers. I present additional supporting evidence for the claim that consumers understand some coarse features of the design in Appendix 4.

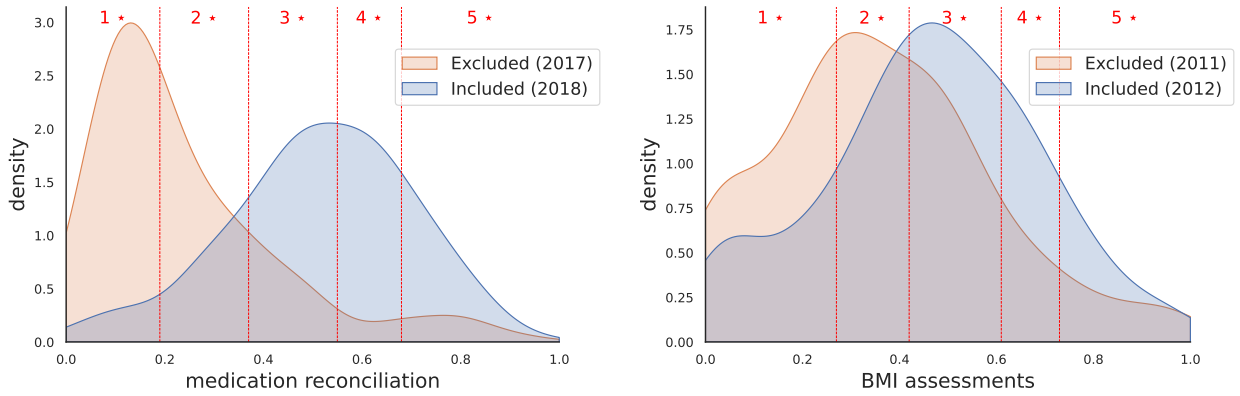
### 4.2 Quality Responses

To explore the causal link between design and quality, I examine variation stemming from the introduction of new quality measurements to the score. For example, in 2018, CMS introduced a Process quality dimension measuring the frequency with which plans updated enrollees' medication records following a hospital discharge. This task prevents medical errors due to changing medication and helps insurers keep track of high-risk drug utilization in their population. Figure 3a illustrates the distribution of this quality before and after the change in design. In 2017, when it was measured but not included in the score, the average plan reconciled medication lists about 20% of the time. In 2018, when the measure was introduced, the average shifted to nearly 60%.

I study eight events of measure introduction occurring between 2009 and 2019.[25] For each, CMS announced the change to insurers without anticipation and measured quality before and after. However, the regulator did not select measures at random, and quality was often already on an improving trend. To isolate the effects of design changes from this trend, I leverage how coarse scores create asymmetric incentives across firms. Plans of low preexisting quality (e.g., those reconciling medication lists 20% of the time in 2017) face a strong incentive to improve as otherwise their scores and demand would decrease. In contrast, plans of high preexisting quality face negligible incentives as the scoring system bins quality in each dimension. This binning stops improvements for high-quality plans from reflecting in the scores, limiting firms' incentives. For example, in Figure 3a a plan that was reconciling medication lists 70% of the time in 2017 would have its improvements pooled within the same fifth bin as if it made no progress. Those bins correspond to the measure-level scores in equation (1).

I use the asymmetric incentives created by the scores to study the effects of design changes on quality. If plans' quality in a measure followed an exogenous trend instead of responding to scoring incentives, it would progress similarly before and after its introduction to the system. If, instead, the shifts in distribution illustrated in Figure 3 are partly due to scoring incentives, then

[25] For the complete list of measures, see Appendix Table 3. There are also events in which measures exit the design, but data on quality following those events are noisy and often missing.




(a) Medication reconciliation

(b) BMI assessments

Figure 3: Changes in quality distributions following introduction to scores

*Notes*: The figures display the distribution of quality (Gaussian kernels) for two dimensions that enter the scoring design during the period of study. The vertical lines indicate how the scores bin the quality distribution in the year the dimensions are included. Each segment corresponds to a measure score. In both figures, the horizontal axis can be interpreted as how often a plan performs the quality process.

high- and low-quality plans should evolve differently after the change in design, as opposed to before it. I use this idea to specify a triple-differences regression that compares the evolution of quality across dimensions, plans, and time:

$$\underbrace{q_{ljt}}_{\substack{\text{normalized quality}\\ \text{measurement}}} = \underbrace{\sum_r \beta_r T_{lt} \mathbb{1}\{G_{lj} = r\}}_{\substack{\text{introduced measure}\\ \text{x quality group}}} + \underbrace{\gamma_{lj} + \mu_{lt} + \xi_{jt}}_{\text{pairwise fixed effects}} + \epsilon_{ljt} \tag{2}$$

Above, $T_{lt}$ indicates if measure $l$ is included in the scoring design at year $t$, and $G_{lj}$ indicates the preexisting quality group of plan $j$ in dimension $l$. I define the groups according to the predicted measure-level score of each plan-measure using the design of the year of introduction but the quality of the preceding year.[26] For example, in Figure 3a, I classify a plan reconciling 40% of its medication into the third group. I normalize the coefficient of interest ($\beta_r$) for the fourth group to zero. Thus, $\beta_r$ captures the change in quality of group $r$ relative to that of the fourth group.

There are three differences involved in the analysis. First is a comparison within plan-measure, controlled by the fixed-effect $\gamma_{lj}$. If only the post-indicator ($T_{lt}$) and this variable were included, then the coefficient on the indicator would reveal if, on average, quality increased following the design change. The second difference, captured by the quality groups ($G_{lj}$) and the measure-time

[26] Formally, $G_{lj} = \text{MeasureScore}_{lt}(q_{ljt-1})$ where $t$ is the year in which measure $l$ is introduced. Appendix 4 presents an alternative analysis using quartiles of the preexisting quality distribution and finds similar results.

fixed-effect $\mu_{lt}$, makes the comparison across groups. In this case, $\beta_r$ would be positive for group $r$ if quality improved more for this group after the design change than for the comparison group. Finally, the third difference compares across dimensions within a plan using the plan-year fixed-effect $\xi_{jt}$. This accounts for the evolution of overall quality in each plan and the trend in MA. Thus the analysis compares the quality change in plan-measures, accounting for the general trends in quality in each dimension and plan. The coefficient of interest is identified from variation in quality within measure, across time, and its differential evolution across quality groups.

I implement the regression by first transforming all quality measurements to a common scale, standardizing them using their mean and standard deviation across all years. Second, I drop plan-measures in the first and last quartiles of quality in the year before introduction to avoid conflating the effects of bounded quality domains.[27] This censoring of preexisting quality drops groups one and five, leaving three levels of $\beta_r$ which capture quality effects relative to the fourth high-quality group. Table 2 shows the results of this analysis.

The first column shows the differences-in-differences specification, which excludes the plan-year fixed-effect. The omission ignores plan trends and overestimates the effect on those with the highest incentive to improve. The second column shows the triple-differences coefficients. The magnitudes decrease with the predicted score as incentives to improve drop. After introduction, a contract-measure that would have obtained a single star under its preexisting quality improves by 0.43 standard deviations more than a contract predicted to get four stars. This is about 40% of the preexisting quality gap between the first and fourth group. Appendix 4 provides further details and supporting evidence for this analysis, showing the lack of pre-trends, effects over time, and robustness to common concerns with staggered differences-in-differences and dynamic treatment effects (Goodman-Bacon, 2021; Baker et al., 2021).

Overall, the exercise reveals that firms respond to scoring incentives by adjusting their quality. Adjustments happen quickly and vary depending on the stakes firms have in responding. These are valuable facts for designing a scoring system, as they inform the extent to which the planner can alter quality in the market through scores. However, as CMS did not select measures at random, the results do not speak to the effect of scoring a generic quality dimension. Others may be more challenging to adjust and less affected by design incentives. The variation needed to measure these effects exists in the data, as scoring rules change every year, yet disentangling them from the overall data variation requires a more structured approach.

The following section presents a model of insurance and demand that rationalizes the reduced-form scoring effects. The model allows me to further leverage MA's extensive design variation to

[27] The domain of most quality measures is bounded. Therefore, low-quality plans can only improve, and high-quality plans can only worsen, and a failure to account for this would inflate the measurements of this analysis. By censoring quality, I likely err on the side of under-estimating the effect.

Table 2: Quality Responses to Scoring

| | Differences-in-Differences | | Triple-Differences | |
|---|---|---|---|---|
| **Predicted score ($\beta_r$)** | | | | |
| 1 | 0.485*** | (0.106) | 0.428*** | (0.108) |
| 2 | 0.365*** | (0.100) | 0.316** | (0.104) |
| 3 | 0.076 | (0.053) | 0.045 | (0.051) |
| Contract measure FE | Yes | | Yes | |
| Measure year FE | Yes | | Yes | |
| Contract year FE | No | | Yes | |
| N | 167693 | | 167693 | |
| $R^2$ | 0.678 | | 0.693 | |
| Mean standardized quality | 0.0319 | | 0.0319 | |

*Notes*: The estimated effect is relative to plans predicted to obtain a measure-level score of 4. The dependent variable is standardized quality in each measure, relative to the mean and standard deviation across all years. To avoid boundary issues, the first and last quartile of pre-treatment quality are excluded. Standard errors, in parentheses, are clustered at the contract level. For further details see Appendix 4.

uncover the primitives that govern these effects. In particular, firms' responses depend on their cost of adjusting quality, and consumers' responses depend on their preferences and beliefs about quality. I use variation in product characteristics and enrollments across markets and time to recover consumers' preferences for insurance plan attributes. Changes in demand and subsidy rules perturb the marginal revenue of insurers, which I use to estimate their marginal cost. Finally, I exploit the evolution of scoring designs to estimate insurers' investment costs and consumers' preferences and beliefs for quality.

# 5 Model

I model the MA demand and supply behavior as the Perfect Bayesian equilibrium of a game consisting of repeated static interactions between consumers and insurers. At the beginning of each year, the regulator announces a national quality scoring rule. Insurers simultaneously choose investments that stochastically determine plans' qualities. They then select prices for their products, which subsidies and regulations convert to premiums and cost-sharing benefits. Finally, consumers observe premiums, cost-sharing benefits, and scores and choose whether to enroll in TM or one of the MA plans available in their county. Figure 4 illustrates the game's timing and information, with bold letters denoting vectors.

Next, I present the model associated with each stage of the game in reverse order. I omit the

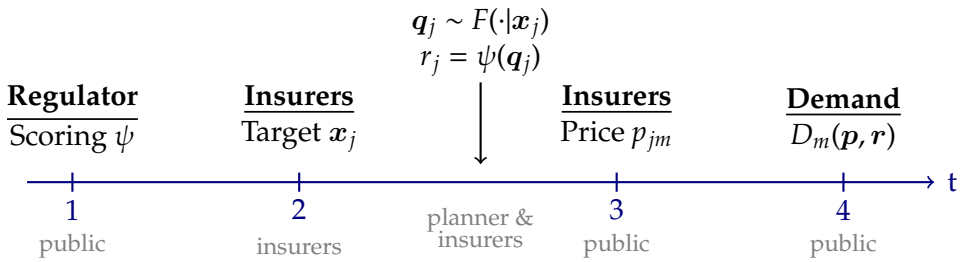



Figure 4: Model Timing

*Notes:* $j$ indexes plans and $m$ markets. $\boldsymbol{x}_j$ denotes an plan's quality investment and $\boldsymbol{q}_j$ its realization. $r_j$ is the score assigned to plan $j$ by the rule $\psi(\cdot)$. The gray text under each stage indicates who observes the choices in the stage.

regulator's choice stage as I do not impose any optimality conditions on the observed scores. Section 7, which solves the scoring design problem, presents their objective. I discuss the implications of the model's central assumptions at the end of this section.

### 5.1 Demand

I model the demand for MA plans following Aizawa and Kim (2018) and Miller et al. (2019), but diverging in two aspects. First, I follow Curto et al. (2021a) and include the dollar value of cost-sharing benefits instead of the different deductibles and coinsurance rates that define it. Consumers observe a similar value in the enrollment platform and CMS regulates cost-sharing at this aggregate level. Second, I model quality preferences explicitly which will become crucial when evaluating counterfactual scoring designs.

Each year $t$, consumers in county $m$ are offered a collection of MA insurance plans $\mathcal{J}_{mt}$. Each plan is characterized by a total premium $p^{\text{total}}_{jmt}$, cost-sharing benefits level $b_{jmt}$, additional plan attributes $\boldsymbol{a}_{jmt}$ (e.g., bundled vision and dental insurance), and a score of $r_{jt}$. Consumers maximize a Von Neumann-Morgenstern expected-utility, evaluating subjective beliefs over quality. The expected utility of consumer $i$ of choosing plan $j$ in market $m$ at year $t$ is given by:

$$u_{ijmt} = \underbrace{\alpha_i p^{\text{total}}_{jmt}}_{\text{premium}} + \underbrace{\beta_i b_{jmt}}_{\text{benefits}} + \underbrace{\mathcal{E}[v(\boldsymbol{q})|r_{jt}, \psi_t]}_{\text{quality}} + \underbrace{\boldsymbol{\lambda}^{a\prime} \boldsymbol{a}_{jmt}}_{\substack{\text{plan} \\ \text{attributes}}} + \underbrace{\boldsymbol{\lambda}^{l\prime} \boldsymbol{l}_{ijt}}_{\substack{\text{lock-in} \\ \text{indicators}}} + \underbrace{\xi_{jmt}}_{\substack{\text{unobserved} \\ \text{preference}}} + \underbrace{\varepsilon_{ijmt}}_{\sim T1EV} \tag{3}$$

Consumers have heterogeneous preferences for premiums and benefits $(\alpha_i, \beta_i)$, and value quality according to the subjective expectation of a common function $(\mathcal{E}[v(\boldsymbol{q})])$ given the plan's score $(r_{jt})$ and the current scoring design $(\psi_t)$. For the time being, I impose no further assumptions on this third term, which captures consumers' preferences for score $r_{jt}$ in year $t$. Enrollment choices are also affect by the plan's bundled services $(\boldsymbol{\lambda}^a)$, and by previous relationships with firms and

products in the market ($\lambda^l$). Following Handel (2013), this last factor captures MA enrollment inertia (Nosal, 2011) as a direct utility impact. Finally, consumers have time-varying unobserved preferences for plan-markets ($\xi_{jmt}$) and an independent random type-1 extreme preference shock perturbs the utility of each choice ($\varepsilon_{ijmt}$).[28]

Consumers also have the option of choosing TM coverage. Given that the vast majority of MA consumers choose plans with bundled prescription drug coverage, I assume that the relevant outside option for the population corresponds to a bundle of TM and stand-alone part D coverage. I denote by $b_0$ TM's standard insurance benefits, and $p^D_{0mt}$ the price of the most popular part D plan in county $m$ and year $t$. Consumers of different demographics might also have specific preferences for TM relative to MA, captured in ($\lambda^d$). The utility of the outside option is given by

$$u_{i0mt} = \underbrace{\alpha_i p^D_{0mt}}_{\text{premium}} + \underbrace{\beta_i b_0}_{\text{benefits}} + \underbrace{\boldsymbol{\lambda}^{d\prime}\boldsymbol{dem}_{it}}_{\substack{\text{consumer}\\ \text{demographics}}} + \varepsilon_{i0mt} \tag{4}$$

Given this model, the expected demand for product $j$ in market $m$ in year $t$ is the sum of the probabilities with which each consumer chooses the product.

$$D_{jmt} = \sum_{i \in \mathcal{I}_m} s_{ijmt} = \sum_{i \in \mathcal{I}_m} \underbrace{\frac{\exp(\delta_{ijmt})}{\exp(\delta_{i0mt}) + \sum_{j' \in \mathcal{J}_{mt}} \exp(\delta_{ij'mt})}}_{\text{individual choice probability}}$$

Where $\delta_{ijmt} = u_{ijmt} - \epsilon_{ijmt}$, is the expected indirect utility of each option.

## 5.2 Supply

***5.2.1 Insurers' pricing problem:*** Each year $t$, at the third stage of the game, insurance firm $f$ observes the vector of realized qualities $\boldsymbol{q}_t$, and its associated scores vector $\boldsymbol{r}_t$. Given this information, the firm chooses prices to maximize its total profits.[29]

$$V_{fmt}(\boldsymbol{q}_t, \psi) = \max_{\{p_{jmt}\}_{j \in \mathcal{J}_{fmt}}} \sum_{j \in \mathcal{J}_{fmt}} \underbrace{D_{jmt}(\boldsymbol{p}_{mt}, \boldsymbol{r}_t)}_{\text{demand}} RA_{jmt} \left( \underbrace{p_{jmt} + R(p_{jmt}, \boldsymbol{z}_{jt})}_{\text{marginal revenue}} - \underbrace{C(\boldsymbol{q}_{jt}, \boldsymbol{a}_{jmt}, \boldsymbol{\theta}^c)}_{\text{marginal cost}} \right) \tag{5}$$

[28] Consumers in MA and TM must pay a fixed part B premium. As this premiums is common across all options, I normalize it in this exposition.

[29] The price of a plan in MA is often called its bid. I avoid this terminology to prevent confusing this market's organization with an auction.

In this equation, each plan's demand is multiplied by the risk-adjustment factor $RA_{jmt}$, which CMS determines for the plan before pricing or demand are realized. The plan's marginal revenue is the sum of its price and additional revenue sources $R(\cdot)$. The second source is, in part, due to the market's regulation and depends on the plan's price ($p_{jmt}$) and attributes ($\boldsymbol{z}_{jt}$). These attributes include its counties of service, prescription drug coverage prices, and the way in which the firm allocates certain subsidies into consumer benefits. I present the full formula for this function and the way in which prices map to premiums and benefits in appendix 5. The cost of covering each enrollee's standard Medicare benefits, prescription drugs, and any non-Medicare extra benefits (e.g., dental insurance), as well as management costs, are contained in $C(\cdot)$. This function varies according to the plan's quality ($\boldsymbol{q}_{jt}$), additional attributes as included in the demand ($\boldsymbol{a}_{jmt}$), and a set of unknown parameters to estimate $\boldsymbol{\theta}^c$, which are the only unknowns of this stage of the model.

The market's price and benefit regulation introduce a kink in the demand and revenue of a firm as a function of prices. If the firm sets prices above the kink, called the plan's benchmark, then a dollar increase in prices translates to an equivalent increase in revenue and premiums, and cost-sharing is not affected. Below the kink, a dollar increase in prices translates into less than a dollar increase in revenue and premiums and a mandatory decrease in the cost-sharing benefits of the plan, in an amount not exceeding a dollar.

***5.2.2 Insurers' investment problem:*** In the second stage of the game, each firm observes the regulator's chosen scoring rule $\psi_t$ and chooses an investment level $x_{ckt}$ for each of its contracts $c$ and category of quality $k$. For example, an insurer can invest in forming networks with better providers to improve its Medical Outcome quality, or hire additional staff to follow up on the well-being of members to improve its Process quality. Firms chooses these investments to maximize their total expected profits.[30]

$$\pi_{ft}(\psi_t) = \max_{\boldsymbol{x}_{ft}} \sum_m \underbrace{\int \mathbb{E}_{mt}[V_{fmt}(\boldsymbol{q}_f, \boldsymbol{q}_{-f}, \psi_t)] dF(\boldsymbol{q}_f|\boldsymbol{x}_{ft})}_{\text{expected insurance profit}} - \underbrace{I(\boldsymbol{x}_{ft}, \boldsymbol{\mu}_{ft})}_{\text{investment cost}} \tag{6}$$

The firm's total expected profits are equal to its expected insurance profit ($V_{fmt}(\cdot)$) in each market $m$ minus the cost of quality investments ($I(\cdot)$). Costs are known functions of each firm's choices and some unknown parameters $\boldsymbol{\mu}_{ft}$.

To derive an expectation of its profits, firm $f$ evaluates two dimensions of uncertainty. First, realized quality might differ from its intended target, captured by the conditional distribution $F(\boldsymbol{q}_f|\boldsymbol{x}_{ft})$. I model the connection between observed qualities and unobserved investments as

[30] Each contract is associated with a set of plans $\mathcal{J}_{ct}$ such that $\mathcal{J}_{ft} = \bigcup_{c \in \mathcal{C}_{ft}} \mathcal{J}_{ct}$.

$q_{ckt} = \Phi_k(x_{ckt} + \epsilon^M_{m(c)kt} + \epsilon^F_{f(c)kt})$, with $\Phi_k(\cdot)$ being a known strictly increasing function. There are two independent (from each other and the target) mean-zero errors in this expression. The first, $\epsilon^M_{m(c)kt}$, captures market-level shocks to a contract's quality and is common across all other contracts offered in the same market. This term captures population distortions to quality supply, such as a harsh flu season or a community vaccination drive. The second, $\epsilon^F_{f(c)kt}$, captures unexpected firm-level deviations in the production of quality, such as cost shocks to provider contract negotiations or firm-level congestion in following up with patients. Together, these shocks capture firms' imperfect control over the quality of their plans, which I document in Appendix 3.1. In practice, insurers form networks and write contracts that attempt to achieve certain targets but might fall short of or exceed their intended goals.

The second dimension of uncertainty is about rivals' investment costs, and therefore, their choices in this stage. As rival investments affect the firm's profits only insofar they shift quality, each firm takes expectations over these realizations ($\boldsymbol{q}_{-f}$). I assume that firms hold rational expectations over the distribution of rival qualities, formed through observation of market characteristics at investment time. These characteristics include the identity of their rivals in each market, the demographic characteristics of consumers, and their previous contract choices. The assumption is motivated by the secrecy of insurers' contractual arrangements with providers and the lack of data sources about quality investments. It is similar to the one made in Sweeting (2009).

The cost parameters ($\boldsymbol{\mu}_{ft}$) and the conditional distribution mapping investments to quality ($F(\boldsymbol{q}|\boldsymbol{x})$) are the two unknowns of this stage of the model. The latter distribution is fully specified as a combination of the unobserved distributions of investments and quality shocks.

### 5.3 Discussion

The model makes two simplifications that might affect the scoring design analysis. First, consumers have homogeneous preferences for quality. This assumption reduces the computational cost of solving the scoring design problem – a stochastic optimization over a non-smooth functional space. My method for solving this otherwise intractable problem suffers from a curse of dimensionality that makes the cost of added heterogeneity gargantuan. A moderate amount would increase the time required to solve this problem from months to years.[31]

Nevertheless, this simplification is unlikely to impact the central question of this paper meaningfully. In Appendix 5.3, I show the model is sufficiently flexible to generate over- and underpro-

[31]The following dimensions are multiplied in the computational cost: dimensions of quality investments, number of rival firms, dimensions of quality shocks, and heterogeneity in consumer preferences. Therefore, in settings with fewer firms or dimensions of quality investments, my method could be used to solve the scoring design problem with heterogeneous quality preferences within a reasonable time.

vision of quality, and thus capture key inefficiencies in quality provision. Moreover, as quality is a vertical attribute, any heterogeneity would consist of enrollees preferring some quality dimensions over others. The alternative design I develop improves quality in almost every category, with those that worsen doing so only moderately. Finally, in Section 8, I show scores can be designed without this assumption or knowledge of consumer preferences, at some loss of optimality.

The second key simplification is that the game is static. Consumers do not learn from their past experiences, and firms do not carry over investments from previous years. Quality in MA, however, is primarily the outcome of contractual arrangements that change often and rapidly. The variation I document in sections 3 and 4 supports this claim. Moreover, the largest insurers in MA have been in existence for decades and likely already invested in major components such as developing relationships with providers or software to track their populations' health. Therefore, dynamic investment incentives are likely of second order in this market. For consumers, the argument in favor of the assumption is similar.[32] The data does not suggest significant differences among MA insurers in their ability to produce quality, which compounded with significant quality variation, makes it improbable that information acquired in a given year will be valuable the next one. Moreover, only consumers with severe health complications will likely learn the more nuanced quality dimensions (e.g., hospital network quality). They are also likely to be the least affected by a change in scoring design due to the switching costs associated with ongoing treatment or illness.

# 6 Identification and Estimation

This section discusses how the data identifies the model presented above and how I estimate its unknown components. I formalize two identification arguments: the non-parametric identification of the distribution mapping investments into quality; and the semi-parametric identification of consumers' quality preferences and beliefs. I relegate the formal statements and proofs to Appendix 6, including technical details about the implementation and estimation steps.

## 6.1 Demand

I estimate the demand model using the two-step approach of Goolsbee and Petrin (2004). The first step uses the individual-level enrollment data to recover preference heterogeneity and uses aggregate market shares to pin down common utility components. It first splits the premium and benefit preferences in equation (3) $(\alpha_i, \beta_i)$ into their mean $(\alpha, \beta)$ and variation $(\tilde{\alpha}_i, \tilde{\beta}_i)$. It then aggregates all common components of the utility of a given enrollment option in a single scalar

[32] The is also a statistical problem. Significant inertia hampers the separate identification of learning from switching costs. The rotating-panel structure of the MCBS further complicates this, as it follows consumers for only a few years.

$\delta_{jmt}$, including mean preferences for premiums, benefits, plan attributes, and quality. This rewriting of the demand model transforms it into one of only five unknown components: preference heterogeneity ($\tilde{\alpha}_i, \tilde{\beta}_i$), demographics preferences for TM ($\boldsymbol{\lambda}^d$), switching costs ($\boldsymbol{\lambda}^l$), and a series of plan-market-year fixed effects ($\boldsymbol{\delta}$). Collecting these components in a vector $\boldsymbol{\vartheta}$, the first-stage solves:

$$\max_{\boldsymbol{\vartheta}} \underbrace{\sum_t \sum_i w_{it} \sum_{j \in \mathcal{J}_{m(i)t}} y_{ijmt} \ln(s_{ijmt}(\boldsymbol{\vartheta}))}_{\text{weighted log-likelihood}} \quad \text{s.t} \quad \underbrace{s^*_{jmt} = \sum_i w_{it} s_{ijmt}(\boldsymbol{\vartheta})}_{\text{share matching}} \quad \forall j, m, t \tag{7}$$

Where $y_{ijmt}$ indicates that consumer $i$ chose plan $j$ in the respective county-year, $s_{ijmt}(\boldsymbol{\vartheta})$ is the model-implied individual choice probability, and $s^*_{jmt}$ is the observed market share.

Overall, the first step consists of a constrained weighted maximum likelihood problem. The weights, $w_{it}$, adjust the MCBS sampling frequencies to represent the national population. The constraint imposes that predicted and observed market share match, which I solve using the Berry (1994) inversion and the Berry et al. (1995) fixed-point contraction.

The second step of the estimator is a two-stage least-squares regression of the estimated mean preference ($\hat{\boldsymbol{\delta}}$) onto its components, recovering all remaining utility parameters. Firms' knowledge of consumers' unobserved preferences for products ($\xi_{jmt}$) when pricing creates a correlation among the second-stage residual, premiums, and benefits. To address this endogeneity, I develop instruments based on regulatory features of insurers' additional revenue ($R(\cdot)$). First, I leverage variation in the kink of $R$ across plans and years. The kink's location depends on TM's cost in every county in which the plan participates, suggesting using TM's cost in every other market in which the plan operates as an instrument.[33] By construction, this instrument is unlikely to express any systematic preference for a specific MA plan as it is associated with the outside option's cost in other markets. Moreover, the second stage includes market and contract-year fixed effects, limiting the residual unobserved preferences.[34] The second instrument uses variation across plans in the added revenue they obtain when pricing below the kink.[35] While only the plan's price is endogenous, this second variable helps distinguish between its effect on premiums and benefits. As the second stage includes year-contract fixed effects to capture quality preferences, the instrument varies only across plans due to county choices. Both instruments have yearly variation caused by changes in regulation and TM's cost. Appendix Table 13 presents the first-stage estimates.

[33]Specifically, the first instrument consists of the leave-one-out average of market-level benchmarks for each plan-market-year, excluding the current market.

[34]Failure of the exclusion restriction would require, for example, plans to change counties as the correlation between TM cost and plan-specific preference varies. As 92% of non-terminated plans remain in a county the following year, this concern seems unlikely.

[35]This instrument corresponds to the rebate fraction for plans pricing above the benchmark and one for the rest.

***6.1.1 Quality beliefs and preferences:*** The previous estimation step recovered consumers' preference for scores ($\mathcal{E}[v(\boldsymbol{q})|r_{jt}, \psi_t]$) as part of a contract-year fixed effects. The model can be estimated without imposing further structure on this component. However, counterfactual changes in the scoring design ($\psi_t$) will affect this object, as consumers will see new scores assigned under new rules. This demands additional structure.

The standard assumption in the literature is that consumers understand the scoring design and derive posterior expectations through the Bayesian updating of some prior belief. For example, in the Bayesian persuasion literature (Kamenica, 2019), signal receivers (consumers) have accurate priors over the state (quality) and understand the signal structure (design). In the empirical literature, several papers have used parametric models of Bayesian demand assuming that signal receivers understand its structure (Crawford and Shum, 2005; Dranove and Sfekas, 2008; Chernew et al., 2008; Brown, 2018; Jin and Vasserman, 2019; Barahona et al., 2020).[36] I rely on a similar assumption for the main counterfactual analysis. As supported by the descriptive evidence, I assume that consumers understand how the scoring system partitions the space of qualities at the category level. To illustrate, Figure 5 shows a scoring rule that uses only three stars and classifies plans according to their average performance in medical outcomes and process measures. The design is represented by lines segmenting the space of qualities, and the assumption is that consumers understand the location and slope of these lines. With added scores, more lines are drawn, and with additional dimensions of quality, lines become hyperplanes. The fact that scores in MA are well represented as partitions stems from CMS's choices of categories and weighting schemes.[37] It is also an inherent feature of partitional scoring designs, such as quality certifications.

This assumption, which I call *informed choice*, requires consumers to know the contribution of categories to the scores and certain cutoff values that determine where one score ends and a new one begins. Variation in category contribution is largely due to changes in the number of measures composing each, shown by CMS in the enrollment platform. Cutoffs depend on measure-level scores that change only moderately from year to year, allowing some learning to occur. In total, the assumption states that consumers' hold some prior $f$ over quality, which they update after observing a score $r$, to compute the expectation of $v(\boldsymbol{q})$ conditional on $\boldsymbol{q}$ being within a set $\boldsymbol{Q}_r$. The assumption does not require the prior to be accurate or imposes any parametric structure on the distribution of signals, relying instead on its true features.

Naturally, consumers' preferences for scores are identified from their willingness to trade premium increases for them, all else equal. Scores are an observable product attribute and standard

[36] Alternatively, some assume that consumers interpret the signal as if it originates from a specific and known parametric signal linked through moments to the data.

[37] Appendix 6.1 shows how the MA scores can be reconstructed at this level with minimal loss.

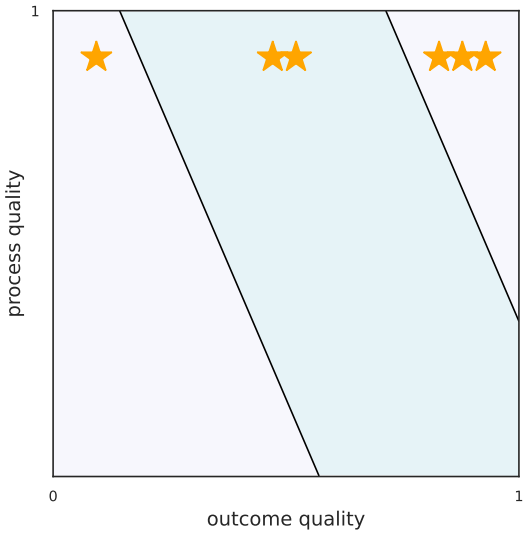



Figure 5: Scores in two dimensions

*Note*: This figure shows an example of scores that partition the space of Process and Outcome quality into three separate signals. The informed choice assumption states that consumers know the location and slope of the scoring lines.

revealed-preference arguments apply. The identification challenge is to tell, for example, whether consumers prefer four-star to three-star plans because they believe the quality difference is small but valuable (i.e., $\gamma$ is large) or because they value marginal quality changes little but believe the difference is large (i.e., $\gamma$ is small). These two configurations might generate the same choice data for a single year, but as I show in Appendix Theorem 1, variation in scoring design will eventually lead them to generate systematically different choices. Specifically, I show that under the assumption of informed choice and that $v(\boldsymbol{q})$ is linear, choice data separately identifies preferences from beliefs, without imposing any parametric restriction on the prior.[38]

The intuition behind this result is that consumers' willingness to pay for score increments implies bounds on their preferences and beliefs. For example, suppose quality is scalar, the prior is uniform, and $v(q) = \gamma q$ with $\gamma = 1$. If there are nine scores uniformly dividing $[0, 1]$, then consumers would be willing to pay $8/9$ more for a top-rated product ($q \in [8/9, 1]$) than for bottom-rated one ($q \in [0, 1/9)$). Some simple algebra shows that by observing the differences in willingness to pay, and knowing the scoring structure, $\gamma$ can be bounded within $(8/9, 8/7)$. Scoring variation produces new intervals for $\gamma$, which intersect and shrink the identified set down to a point. This structure also bounds posterior beliefs, and thus priors. This example and its formal counterpart rely on the identification of consumers' valuation for score-years. Berry and Haile (2020) provide conditions

[38] An alternative, fully non-parametric argument, would note that scores define lotteries over qualities and appeal to the result of Anscombe and Aumann (1963). However, this argument would require scores to cover the entire space of lotteries, which is unrealistic for the limited type of designs in MA. The result I derive only uses variation consistent with the data.

for identifying such systematic preferences from individual-level choice data. Importantly, thanks to their result, my proof does not rely on the logit structure.

The result shows that informed choice is a powerful assumption. It imposes a strong structure on consumers' understanding and, in return, delivers identification. However, it is not strictly necessary in order to design scores. In Section 8, I assume instead that consumers are entirely uninformed of changes to the scoring design. In the appendix, I show that this assumption implies that preferences are only set-identified, proving that lacking assumptions on how consumers interpret scores, choice data alone does not identify preferences and beliefs over quality. This observation might explain why successful disclosure systems are often accompanied by a public information campaign, as consumers' understanding of the system is helpful to them and the regulator (Barahona et al., 2020). However, as I will discuss, the worst-case scenario for consumers' preferences might still be bounded and used to design a system that improves welfare.

For the results that rely on the informed choice assumption, I estimate preferences (now captured by a vector $\gamma$), and prior beliefs ($f(\cdot)$) using a non-parametric minimum distance estimator. To remove any systematic preference for specific contracts, I only leverage time-series variation within the contract's valuation, $\eta_{c(j)t} \equiv \mathcal{E}[\boldsymbol{q}|r_{c(j)t}, \psi_t]$, which I recover in the second stage of the main demand estimator. The resulting estimator is

$$\min_{\gamma,\zeta} \sum_{c(j)} \sum_{t} \sum_{\tau>t} \left(\Delta_t^\tau(\eta_{c(j)t} - \gamma' \mathcal{E}[\boldsymbol{q}|r_{c(j)t}, \psi_t; \zeta])\right)^2 \tag{8}$$

where $\Delta_t^\tau x_t \equiv x_\tau - x_t$ is the time difference operator and $\zeta$ corresponds to the Fourier-coefficients of a series expansion of the common prior $\boldsymbol{f}(\cdot)$ onto a Fourier series. This step does not affect other estimates and can be safely disregarded when relying on the assumption of ignorance.

***6.1.2 Estimates:*** Table 3 presents the main demand estimates. Panel A shows the estimated preferences for premium and benefit levels. A dollar in benefits is roughly equivalent to a two-and-a-half dollar reduction in premiums for a low-income male of "fair" perceived health. Poorer and healthier consumers are more responsive to premiums and benefits. The distaste for premiums decreases with age but benefits preferences are concave, peaking between 70 and 75. The average price elasticity – a statistic that aggregates premium and benefits preferences – is -8.34. This is the elasticity relevant for firms' pricing decisions and, in particular, a single-product monopolist with constant marginal cost and no part D coverage would set prices to meet an elasticity of -1.[39] As I

[39]The statement about the monopolist assumes it would price above the benchmark, as is often the case in the data. The table also display premium elasticities which can be compared to those of Miller et al. (2019). Their estimates is of -2.6 using similar data but a different model. Such a high premium elasticity would imply excessive price elasticities leading to negligible firm markups in my model. Appendix Figure 3 display the estimated price elasticities.

will show later, this elasticity implies reasonable markups for firms in this market.

Panel B presents some of the additional preferences for fixed product attributes. Importantly, consumers have a strong preference for products bundling prescription drug coverage. Panel C of Table 3 presents consumers' quality preferences under the assumption of informed choice. The most valued category is Medical Outcomes and the least is Intermediate Medical Outcomes. Consumers are willing to pay \$4498 in yearly premiums for maximal Outcome quality and \$1654 for maximal Intermediate quality. However, quality dispersion differs across categories and a standard deviation increase is worth \$204 in Outcomes and \$194 in Intermediate. Appendix Section 6.2 presents the full set of estimated coefficients.

Following Train (2015), I compute the surplus loss from consumers' incomplete information about quality, holding product attributes fixed. I estimate that the average consumer loses approximately \$185.9 per year relative to full information due to two reasons. First, they cannot distinguish between heterogeneous products receiving the same score. The average consumer, if informed, would be willing to pay \$257.1 more for the highest quality four-star plan in its choice set, than for the lowest quality plan of the same score, all else equal. The second reason for the surplus loss is that, on average, 22.4% of plans in a county have a lower-scoring alternative that delivers higher quality-utility. The failure stems from the misalignment between consumers' preferences across categories and the weights assigned by CMS in the scoring design. Thus, although both consumers and CMS agree that plans' of higher overall quality should receive higher scores, they disagree on how to weigh different dimensions. In total, the surplus loss equals three months of average MA premiums (Part C + D).[40]

## 6.2 Supply

I estimate the different components governing price and quality competition in three steps. First, I use optimality conditions from the pricing problem to estimate the marginal cost of insurance. Second, I use a high-dimensional non-parametric conditional density estimator to recover the quality shock distribution. Finally, I compute the expected marginal profits from quality investment to estimate the investment cost. I discuss each in order.

***6.2.1 Insurance marginal costs:*** The first-order optimality conditions (FOC) of the insurer's pricing problem equate marginal revenue with marginal costs.[41] As revenue depends only on observed demands, prices, and estimated elasticities, the firm's FOC can be used to recover the

[40] See Appendix 6.2.2 for details on this analysis.

[41] As the firm's problem is not differentiable at the benchmark, the FOC is only valid for prices away from this cutoff. However, as in the data no firm violates this condition, it holds for all observations.

Table 3: Key Demand Estimates

| **Panel A: Premium and benefit mean preference and heterogeneity** | | | | |
|---|---|---|---|---|
| | **Premium ($\alpha_i$)** | | **Benefits ($\beta_i$)** | |
| Mean preference | -1.112** | (0.393) | 2.915*** | (0.383) |
| Medium income | 0.041 | (0.057) | -0.028 | (0.071) |
| High income | 0.271*** | (0.060) | -0.167* | (0.073) |
| Female | -0.057 | (0.046) | -0.006 | (0.058) |
| Age group $<65$ | -0.111 | (0.091) | -0.005 | (0.099) |
| Age group $\in [70,75)$ | -0.009 | (0.057) | 0.151*** | (0.041) |
| Age group $\in [75,85)$ | 0.017 | (0.055) | 0.102* | (0.040) |
| Age group $\geq 85$ | 0.195* | (0.081) | -0.110 | (0.058) |
| Health - Excellent | -0.262*** | (0.077) | 0.010 | (0.057) |
| Health - Very Good | -0.226** | (0.069) | -0.022 | (0.052) |
| Health - Good | -0.132 | (0.068) | -0.044 | (0.050) |
| Health - Poor | -0.005 | (0.116) | -0.145 | (0.083) |

| **Panel B: Other product attributes ($\lambda^a$)** | | | **Panel C: Quality preferences ($\gamma$)** | | |
|---|---|---|---|---|---|
| Drug deductible | -0.001*** | (0.000) | Access | 4.501*** | (0.365) |
| Part D coverage | 1.778*** | (0.020) | Intermediate | 1.839*** | (0.042) |
| Dental cleaning | 1.846*** | (0.060) | Outcome | 5.002*** | (0.807) |
| Hearing aids | -0.229*** | (0.031) | Patient | 3.792*** | (1.112) |
| Vision insurance | -0.032 | (0.023) | Process | 2.315*** | (0.161) |

| **Panel D: General information** | | | |
|---|---|---|---|
| Observations | 36447 | Weighted log. likelihood | -5.131 |
| Mean price elasticity | -8.348 | Mean premium elasticity ($p^C > 0$) | -0.951 |

*Notes*: Panel A and B report key estimates of individual preference for product attributes corresponding to equation (3). In Panel A, the omitted category is low income males of "fair" self-reported health-status. Income groups are defined by terciles of the MCBS income distribution. Premiums and benefits are measured in thousands of dollars per year. Panel C reports quality preference estimates under the assumption of informed choice. All observations are weighted by the MCBS sample weights. Un-adjusted heteroskedastic standard errors in parenthesis. *p<0.05, **p<0.01, ***p<0.001.

marginal cost parameters ($\boldsymbol{\theta}^c$). Assuming marginal costs are linear, the resulting condition is:

$$\underbrace{\boldsymbol{p}_f + R(\boldsymbol{p}_f, \boldsymbol{z}_f)}_{\text{revenue per consumer}} \underbrace{+ (\nabla \tilde{\boldsymbol{D}}_f')^{-1}(I + \nabla R_f(\boldsymbol{p}_f, \boldsymbol{z}_f))\tilde{\boldsymbol{D}}_f}_{-\text{profit margin}} = \underbrace{\boldsymbol{\theta}_q^{c\prime}\boldsymbol{q}_f + \boldsymbol{\theta}_a^{c\prime}\boldsymbol{a}_f + \boldsymbol{c}_f}_{\text{marginal cost} = \boldsymbol{C}(\boldsymbol{q}_f, \boldsymbol{a}_f, \boldsymbol{\theta}^c)} \tag{9}$$

where gradients are all with respect to the vector of prices $\boldsymbol{p}_f$, and $\tilde{\boldsymbol{D}}_f$ is the risk-adjusted demand vector. The identity states that revenue per consumer minus the firm's profit margin equates the marginal cost. The margin depends on consumers price elasticity and the change in additional revenue produced by CMS's price regulation. On the right-hand side, I have decomposed the firm's marginal cost into its quality components ($\boldsymbol{q}_f$), its systematic observable components ($\boldsymbol{a}_f$),

Table 4: Quality's Insurance and Investment Costs

| | Panel A: Insurance Cost ($\theta_q^c$) | | Panel B: Investment Cost ($\mu_k$) | |
|---|---|---|---|---|
| Access | 31.160 | (16.690) | 15.620** | (5.965) |
| Intermediate | 108.400*** | (12.800) | 19.530*** | (4.963) |
| Outcome | 16.810*** | (3.832) | 15.000* | (6.516) |
| Patient | -244.300*** | (57.540) | 14.730* | (7.424) |
| Process | -175.600*** | (27.560) | 1.106 | (4.718) |
| N | 28966 | | 5281 | |
| R2 | 0.531 | | 0.261 | |

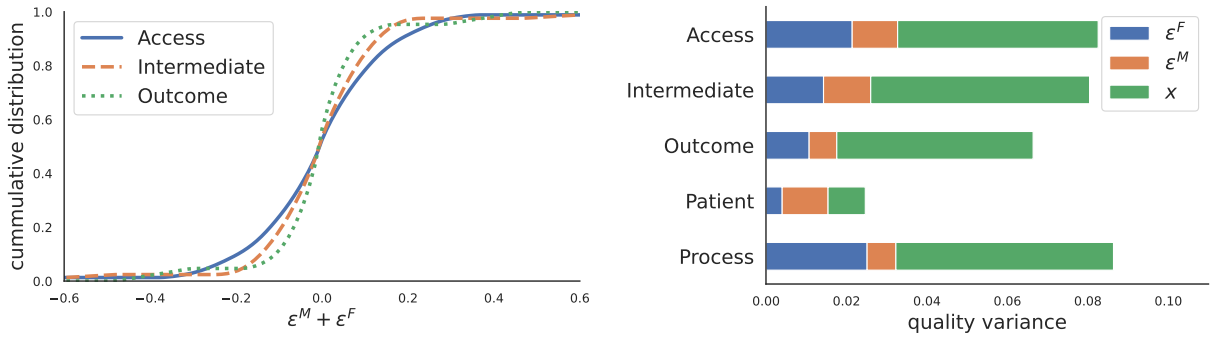
*Notes*: This table reports the estimates of $\theta_q^c$ in the marginal cost equation (9), and $\mu_k$ in the investment cost equation (10). Values on the left are in dollars per member-month, while on the right are in millions per contract-year. Standard errors in parenthesis are heteroskedasticity robust. For further details on the marginal cost see Section 6.2.1. For further details on the investment cost Section 6.2.3. *p<0.05, **p<0.01, ***p<0.001.

and its residual plan-market-year specific component ($c_f$).

Variation in demand, competition, and regulation all serve to identify marginal costs. Panel A of Table 4 presents the estimates of $\theta_q^c$ when $a_f$ includes contract, year, and market fixed effects, as well as controls for all bundled services provided by the plan. The effect of quality on marginal costs is identified by how a plan's marginal revenue varies when its quality changes in ways unrelated to the market or national quality trends. The estimates indicate that improving both types of medical outcomes increases marginal cost. In contrast, improvements in Process and Patient quality lower marginal cost. One justification for this reversal is that process measures include preventive care and the management of expensive chronic illnesses. For an elderly population, these can help prevent expensive hospitalization and reduce costs (Newhouse and McGuire, 2014). Having better physicians in the network – captured by Patient quality – is likely associated with similar improvements and might make patients more likely to adhere to preventive and diagnostic care. These negative costs do not imply that firms should set these qualities to their maximum, as there might still be significant investment costs.

These estimates imply reasonable markups for insurers, with the average being 11.2%, while for the top 4 insurers, it is 13.3%. As a point of comparison, Curto et al. (2019) use the Health Care Cost Institute data to estimate that in 2010, the average insurer in their sample spent $590 per enrollee risk-month in medical costs, or $680 in adjusted 2015 dollars. My estimate for the same set of firms is an average of $771, including medical and administrative costs. This comparison suggests that about 10% of marginal cost is administrative, which is consistent with the level of involvement of MA insurers with their enrollee's health.[42]

[42] Appendix Figure 2 shows the distribution of markups and marginal costs.




(a) Shock distributions

(b) Variance decomposition

Figure 6: Quality shock distribution

*Notes*: These figures display the estimated distribution of quality shocks. Figure (a) shows the cumulative distribution function of the sum of shocks for each category. Patient and Process categories have been excluded for clarity. Panel (b) shows the fraction of the variance of the observed quality that is attributed to investments and the two types of shocks. Additional plots are provided in Appendix Section 6.3.1.

***6.2.2 Quality shocks:*** Because quality investments are unobserved, I cannot recover investment costs by leveraging optimality conditions directly, as done above for marginal costs. Instead, I follow an approach that uses the expectation of firms' FOCs conditional on the observed quality. To do so, I must first uncover the distribution of quality shocks contaminating the data. Using results for the non-parametric measurement error literature (Schennach, 2016), I show in Appendix Theorem 4 that the distribution of quality shock is identified from quality data.

The intuition behind this result is the following. Consider two firms that offer products in two markets. The quality provided by each product is a combination of unknown investments distorted by population and firm-level shocks. As firms' beliefs about rivals depend on market characteristics, investment choices conditional on these characteristics are independent of rival investments. Consequently, any residual correlation in quality across firms within a market is driven by market-level quality shocks. Conversely, any correlation across markets within a firm is due to firm-level shocks. The formal result transforms this intuition into a conditional deconvolution of observable quality distributions which I estimate using the high-dimensional estimator of Izbicki and Lee (2017).[43] In estimation, I take $\Phi_k(x) = \Phi(x)(1-\underline{q}_x) + \underline{q}_k$ where $\Phi(\cdot)$ is the standard normal CDF, and $\underline{q}_k$ is the minimum value of quality $k$ that firms can produce.[44] Appendix 6.3.1 details this estimation step which acts as a prerequisite for investment cost estimation.[45]

In total, this estimation recovers the ten distributions modeled: two shock types for five

[43]Similar results have been used in the auction heterogeneity literature (Krasnokutskaya, 2011).

[44]While the domain of quality is the unit interval, in practice there are minimum standards. For example, an insurer cannot contract with a hospital to act in a way that would actively harm patients.

[45]This estimator does not build on previous estimates. Because of this, and to offset the slower convergence rate of this class of non-parametric estimators (Horowitz and Markatou, 1996), I use the full 2009-2019 data for this estimation.

categories, illustrated in Figure 6. The results show that shocks account for 39.6% of the variation in observed quality. Patients' assessments are the noisiest, as insurers cannot contract for better reviews and, correspondingly, most of the variance is due to market-level shocks. Firm-level shocks are most important in Process measures, as they are insurer-labor intensive, involving following up with patients and helping them schedule appointments for testing and care.

***6.2.3 Investment costs:*** I estimate firms' investment cost parameters ($\boldsymbol{\mu}_{ft}$) by combining the optimality conditions associated with the investment problem and the estimated distribution of quality shocks. Specifically, firms' investments in each dimension $k$ for each contract $c$ maximizes their profits, hence equating marginal insurance profits with marginal investment costs:

$$\frac{\partial}{\partial x_{ckt}}\mathbb{E}[V_f(\boldsymbol{q}_f, \boldsymbol{q}_{-f}, \psi_t)|\boldsymbol{x}_{ft}] = \frac{\partial I(\boldsymbol{x}_{ft}, \mu_{ft})}{\partial x_{ckt}}$$

Where I have slightly compacted the notation of equation (6).

As investment are unobserved, I cannot evaluate the expression above. However, as I observe the distribution of realized qualities and have estimated that of quality shocks, I can compute the distribution of optimal investments overall. That is, I can evaluate the likelihood with which firm $f$, chose $x_{ckt}$ as its optimal investment given that I observe $q_{ckt}$ in the data. Therefore, decomposing the marginal insurance profit into its mean and variation, the FOC can be written as

$$\mathbb{E}[\frac{\partial}{\partial x_{ckt}}\mathbb{E}[V_f(\boldsymbol{q}_f, \boldsymbol{q}_{-f}, \psi_t)|\boldsymbol{x}_{ft}]|\boldsymbol{q}_{ft}] = \frac{\partial I(\boldsymbol{x}_{ft}, \mu_{ft})}{\partial x_{ckt}} + \nu_{ckt} \qquad \mathbb{E}[\nu_{ckt}|\boldsymbol{q}_{ft}] = 0$$

Appendix Proposition 3 shows that the left-hand side of this expression is a function of only identified distributions and gives an analytic expression for it.

I assume firms' investment costs are quadratic and separable across products and categories

$$I(\boldsymbol{x}_{ft}, \boldsymbol{\mu}_{ft}) = \sum_{c \in C_f}\sum_{k}\left(\mu_k(x_{ckt} - \underline{x}_{kt})^2 + \mu^F_{fkt}(x_{ckt} - \underline{x}_{kt}))\right)$$

Where $\underline{x}_{kt}$ is the state-category baseline investment, representing the lowest level of investment a firm can deliver to participate in a state. Anything above this level requires either forming a network or writing contracts to promote quality. Using this expression and decomposing the marginal investment cost into its conditional mean and variance results in the regression equation:

$$\underbrace{\mathbb{E}[\frac{\partial}{\partial x_{ckt}}\mathbb{E}[V_f(\boldsymbol{q}_f, \boldsymbol{q}_{-f}, \psi_t)|\boldsymbol{x}_{ft}]|\boldsymbol{q}_{ft}]}_{\substack{\text{conditional expectation} \\ \text{of marginal insurance profits}}} = \underbrace{2\mu_k(\Phi_k^{-1}(q_{ckt}) - \Phi_k^{-1}(\underline{q}_{kt})) + \mu^F_{f(c)kt}}_{\substack{\text{conditional expectation} \\ \text{of marginal investment cost}}} + \tilde{\nu}_{ckt} \tag{10}$$

The residual component, $\tilde{\nu}_{ckt}$, is the sum of two errors. First, it contains those introduced by substituting marginal profits and costs with their expected values conditional on realized qualities. By construction, this term has a conditional mean of zero. Second, the residual contains the error added by replacing the baseline investment with their closest empirical analog: the minimum quality in the state-year mapped to the space of investments. Assuming that this second error is mean-zero conditional on $\boldsymbol{q}_{ckt}$, equation (10) is a linear regression.

I estimate the marginal investment cost using OLS. To avoid mixing contracts with very distinct cost structures, I limit attention to HMO and PPO contracts, which account for 81% of enrollment. This restriction excludes Private Fee-For-Service contracts, which do not form networks directly, and Regional PPO contracts, which have broad networks that often cross multiple state lines. Second, I separate the problem across states for each firm. While 18% of beneficiaries choose contracts offered in multiple states, the median multi-state contract has 80% of its population in a single state. Panel B of Table 4 displays the estimated coefficients. It shows that Intermediate Outcome quality is the most expensive to improve, while process measures are the cheapest. In comparative terms, the estimates suggest that contracts in the 75th percentile of Access quality invested an average of 5.16 million dollars more than contracts in the 25th percentile. Overall, the median quality firm invests 24.6% of its profits.

## 7 Scoring Design

The previous sections documented that scores shift demand and investment choices. They have also specified a model for the market and recovered estimates of the primitives governing product choice and quality production. Despite its simplifications, the model delivers reasonable estimates. Demand elasticities imply markups that agree with external evidence, and investment spending is moderate but meaningful. This section leverages the previous findings and the assumption that consumers understand the scoring design (informed choice) to specify and solve the designer's problem. The following section studies scoring design without informed choice.

I assume that the designer seeks to maximize the expected weighted total welfare given by:

$$\max_{\psi \in \Psi} \int \left( \underbrace{CS(\psi, \boldsymbol{q})}_{\substack{\text{Consumer} \\ \text{surplus}}} + \rho^F \underbrace{\sum_f V_f(\psi, \boldsymbol{q}) - I(\boldsymbol{x}_f^*(\psi), \mu_f)}_{\substack{\text{Insurer} \\ \text{profit}}} - \underbrace{\rho^G Gov(\psi, \boldsymbol{q})}_{\substack{\text{Government} \\ \text{spending}}} \right) dF(\boldsymbol{q} | \boldsymbol{x}^*(\psi)) \tag{11}$$

The objective sums consumers' surplus and insurers' profits, subtracting the government's spend-

ing on subsidizing enrollees in MA relative to the cost of insuring them under TM. To maximize welfare, the designer publicly announces and commits to a deterministic scoring rule ($\psi$) that partitions the space of quality into distinct scores, assigning greater scores to higher quality. I focus on this class of deterministic monotone designs as they are common, incorporate the MA system, and are optimal under certain scenarios.[46]

The designer faces a trade-off between eliminating information frictions and regulating quality. For any fixed investment level, more information helps consumers choose and makes competition more effective. That is, in general, the integrand in the designer's objective is maximized at full information. The key tension comes from the Spencian distortion: firms' investments are often inefficient. As shown in Section 2, this might lead the designer to choose a coarse scoring rule that trades information for efficiency.

The same tension between information and regulation appears within the consumer surplus component. The designer is concerned with the ex-post surplus generated by matching consumers with products of a certain quality. In contrast, consumers choose based on beliefs about quality given scores, not knowing if the effective match is optimal. Thus consumers' surplus can be decomposed as (Train, 2015):

$$CS(\psi, \boldsymbol{q}) = \underbrace{CS_0(\psi, \boldsymbol{q}) + \zeta}_{\substack{\text{ex-ante} \\ \text{surplus}}} + \underbrace{\sum_{j \in \mathcal{J}} \gamma'(\boldsymbol{q}_j - \mathcal{E}[\boldsymbol{q}_j | \psi(\boldsymbol{q}_j)])(\int \frac{s_{ij}(\psi(\boldsymbol{q}))}{|\alpha_i|} di)}_{\text{Ex-post correction}} \tag{12}$$

The first term is the ex-ante logit surplus, with $\zeta$ its unknown location (Small and Rosen, 1981). The second term adjusts surplus for the gap between consumers' beliefs about quality and its actual value. The trade-off for the designer is that more information shrinks the ex-post correction and increases the ex-ante surplus as consumers match better with products. However, more information might decrease the efficiency of quality, hence lowering the overall surplus.[47]

The remainder of this section is organized as follows. First, I discuss why solving the designer's problem is challenging and how I address it. Second, I present the *best linear substitute* for the MA Stars: a welfare-improving alternative policy that uses the same scoring technology as the incumbent system. Third, I compare this design to a full-information counterfactual, revealing that the unconstrained optimal scores are coarse. Fourth, I study quality certifications, showing

[46] All deterministic quality certifications fall within this class (crash test, organic labels, high-in-sugar food labels, etc.). See Dworczak and Martini (2019) for proof of optimality for similarly defined scores under exogenous quality. Their definition allows some segments to be revealing, which I explore in the appendix.

[47] The designer evaluates consumers' surplus in expectation. Hence the ex-post correction would be zero if consumers were as informed about the distribution of quality as the regulator. In the analysis, however, consumers' prior are independent of the scoring rule to reflect their lack of knowledge about the complexity of firms' behavior.

that simple policies can be remarkably effective if well-designed. Finally, I discuss the effects of incorporating the regulator's private preferences for quality, if they exist, in the objective.

The following results are for a subset of markets in 2015, covering nearly 22 million beneficiaries and for the case of $\rho^F = 1$ and $\rho^G = 0$.[48] I show results for other objectives in Appendix 7. The same includes more sophisticated designs that minimally improve welfare and results regarding the losses from geographic aggregation (i.e., national scores vs. local designs) and the effects of competition on scoring design.

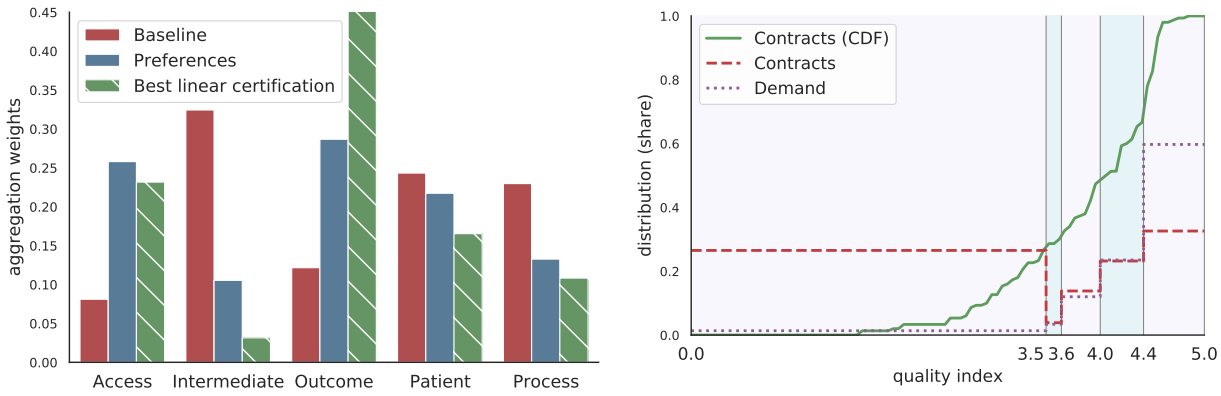
### 7.1 Solution approach

The designer's problem consists of finding a policy to maximize expected welfare. As scores are discrete and summarize multiple quality dimensions, the designer's solution is a discontinuous mapping. There are no known optimality conditions for this problem, and a priori, the loss from any simple approximation might be unbounded. Moreover, as the regulator computes an expectation over quality, evaluating any design requires integrating over counterfactual equilibria.

My solution is a *divide and conquer* strategy. First, I show in Appendix Proposition 4 that any finite monotone deterministic score is a composition of an *aggregator* and a *cutoff* function. The aggregator summarizes multidimensional quality into an index which the cutoff function then segments into scores. The proposition shows that the aggregator is, without loss, a polynomial function. Thus, conditional on the maximum number of scores in a system, the designer's problem boils down to choosing a series of polynomial coefficients and the gap between one score and the next. This problem is well-behaved and can be tackled computationally. However, it is still the case that to evaluate the welfare of any given design, we must compute thousands of counterfactual equilibria. To tackle this challenge, I draw on the insight of Kamenica and Gentzkow (2011), that choosing a disclosure policy is akin to selecting a distribution over posterior beliefs. In the case of scoring design, the analogous statement is that each score generates a distribution over qualities, score valuations ($\mathcal{E}[\boldsymbol{q}|r,\psi]$), and marginal quality costs ($\boldsymbol{\theta}^{c\prime}\boldsymbol{q}$). This observation enables a strategy that first evaluates the objective over a large collection of potential outcomes and then associates each score with a distribution over the collection.[49] Overall, my solution approach takes an intractable problem and divides it into many smaller ones that can be solved independently and then combined. Appendix 7.1 provides further details.

[48] The subset is given by the set of counties covered by the MCBS after the HMO/PPO restriction.

[49] These spaces are compact given that the space of quality is compact.




(a) Aggregation weights

(b) Quality distribution and cutoffs

Figure 7: Best Linear Substitute Design

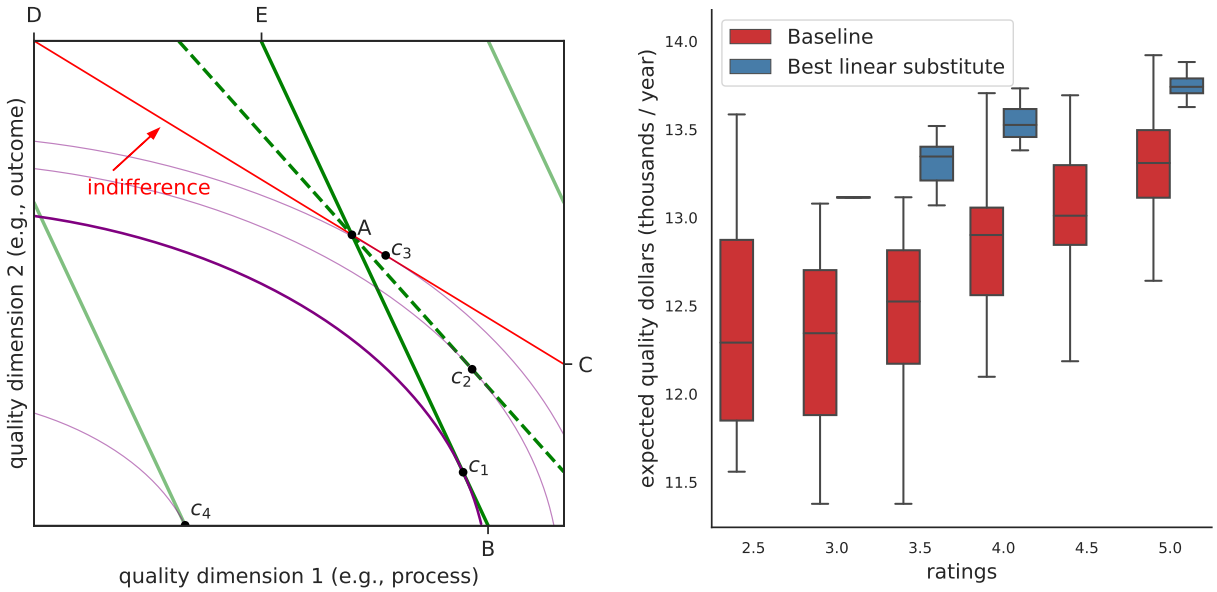
*Notes*: Figure (a) shows the weight on each category by the new aggregator compared to CMS's scheme and consumers' preferences. The green striped bar marks the new design. Figure (b) shows the cutoff placement for new design. The green solid line shows the cumulative distribution of contracts over the index quality, as computed by the aggregator. Vertical segments of changing colors indicate different scores.

## 7.2 Best linear substitute

The best linear substitute design maximizes total welfare using the same technology as the MA Stars. It classifies quality into at most nine distinct scores (i.e., one to five stars in half increments) and uses a linear mapping to aggregate quality dimensions (e.g., a weighted average). To derive this design, I decompose the scores into a linear aggregator and a cutoff function and find their optimal values. Figure 7 displays the resulting design.

The first noticeable difference between the new and incumbent systems is in their aggregators, shown in Figure 7a. The new design removes weight from the Intermediate Outcome category and shifts it towards Access and Outcome. On the firm side, this change affects how investments get distributed across categories. Each firm can be viewed as first selecting an expected score for its product and then finding the cost-minimizing way of reaching it. Hence, conditional on their target, the firm allocates investments across categories independently of consumers' preferences. This idea is illustrated by Figure 8a in a scenario without investment risk. The firm targets the third score and invests at a point where its iso-cost curve is tangent to the scoring threshold ($c_1$). For the designer, firms' disregard for consumers' preferences over investment allocations translates into a multitasking moral hazard problem (Holmstrom and Milgrom, 1991). The designer would like firms to invest considering consumers' preferences, yet the scores hide relative allocations, eliminating firms' incentives to act accordingly.

The new design addresses this problem by improving the alignment of scoring thresholds




(a) Allignment illustration

(b) Quality overlap per score

Figure 8: Effects of aligning category weights with consumers' preferences

*Notes*: Figure (a) illustrates the theory with two quality dimensions and four scores. The line *EB* is the second scoring cutoff, and *DC* is consumers' indifference curve. Products above *EB* obtain a score of three, and below it, two. The misclassification region is *DEA* + *ABC*, as consumers prefer products in the former over the latter. The green dashed line represents the new design. The purple concave lines represent a firm's iso-cost curve such that in the baseline, it chooses quality $c_1$ and gets three stars. It chooses $c_2$ under the new design but would choose $c_4$ under a perfectly aligned rule as the cost of $c_3$ exceeds the gain in demand. Figure (b) shows the ranges of quality utilities binned in each score in the baseline and the best linear substitute design. The baseline has extensive overlap between scores, implying large misclassification areas. The best linear substitute leads to negligible misclassification despite being imperfectly aligned. This design uses fewer scores hence the missing levels.

and consumers' indifference curves. In Figure 8a, the change is illustrated as a tilt in the second threshold (EB line) towards the indifference line (DC), resulting in the new dashed threshold. As the alignment of scores and preferences improves, any substitution along the scoring frontier leaves consumers nearly indifferent. However, the optimal alignment is imperfect as the regulator must consider the effect on firms' investment decisions. Improving the alignment while keeping the cutoff fixed means the firm has to increase its investment from $c_1$ to $c_2$ to maintain its score. While the demand for the product might improve with better alignment, as I will discuss next, the change might be insufficient to offset the convexity of investment costs. Therefore, a perfectly aligned system might make the firm change its target score to a lower one, investing in $c_4$ instead of $c_3$. As this investment is smaller, substantial welfare might be lost.

Aligning the scoring threshold with consumers' quality preferences also alters the informational content of the scores through two channels. First, increasing the weight of a category augments

the correlation between quality in this category and overall scores. Therefore, the new scores are more informative of variation in Outcomes at the expense of less information about Intermediate Outcomes. The second channel is through a decrease in the probability of ex-post mistakes in enrollment choices. Figure 8a illustrates that the misalignment of scoring thresholds and indifference curves creates regions of *misclassification*. Plans falling in the triangle *DEA* receive a lower score than those in the triangle *ABC*, but consumers prefer the former over the latter. Narrowing the regions of misclassification has a non-trivial effect on this market. Figure 8b shows that the MA Stars have a substantial overlap in expected quality utility between scores. This overlap makes it harder for consumers to choose among products and makes the scores less useful for them, in turn eroding the score's ability to marshal demand and thus affect quality.

The second design choice is the cutoff placement. Figure 7b shows that the new design uses only five of its nine available scores, pooling a wide range of low qualities into a single low-quality score.[50] Approximately, the new design would eliminate every star rating below three, giving them a single score. It would narrow the set of plans getting 3.5 and four stars, and leave the top scores of roughly the same width as they currently are. This comparison is only approximate, because the aggregate index is different under the new design. By pooling a wide array of low qualities, the designer creates an effect similar to the one observed in the theoretical analysis of Section 2. Consumers' penalize low-quality products because they cannot tell how low their quality is. They shift their demand towards greater and more informative scores, which induces firms to invest. Overall, only 26.5% of contracts fall in the low-quality segment, and serve but 1.3% of the demand. The highest score contains 32.6% of products and serves 59.7% of consumers.

Table 5 shows the estimated welfare gains from replacing the MA Stars with the best linear substitute. Per Medicare beneficiary-year, the alternative increases surplus by \$146.5, and profits by \$522.7, while slightly decreasing government spending. Consumers gain an equivalent to 2.4 months of average premiums (part C + D) in the baseline, and the additional profits correspond to a 140% increase. Firms profit grow due to 63% higher margins per MA enrollee and a significant market expansion. Consumers switch from TM to MA as quality and information improved, allowing them to select products of higher quality and benefit from MA's more generous cost-sharing. Hence, I interpret this finding as indicating that a large fraction of consumers choose TM because the MA's narrower networks and managed services have uncertain quality. In the counterfactual, consumers substitute often to plans that cost less to subsidize than TM, leading to a decrease in spending of 0.8% per Medicare beneficiary.

To unpack the channels through which the welfare gains occur, I gradually incorporate the counterfactual equilibrium's information, quality, and prices. Table 6 shows the result when eval-

[50] This does not imply that five stars is the unconstrained optimal number of scores, only that it is the best lower or equal to nine. In the appendix I show a design that delivers higher welfare with fifteen scores.

Table 5: Welfare changes in redesigned system under informed choice

| | **Linear** | **Full** | **Certification** | | |
|---|---|---|---|---|---|
| | **Substitute** | **Information** | Best linear | CMS | Preferences |
| Δ Consumer Surplus | 146.5 | 136.0 | 151.8 | 68.6 | 122.8 |
| Δ Firm Profits | 522.7 | 488.6 | 480.5 | 169.1 | 486.5 |
| Δ Gov. Spending | -76.4 | -69.1 | -88.4 | -85.3 | -94.0 |
| Δ Total Welfare | 669.3 | 624.6 | 632.4 | 237.7 | 609.2 |
| MA share | 57.1% | 54.4% | 56.0% | 41.3% | 54.9% |

*Notes*: This table displays the welfare effect of the alternative system, relative to the MA Star Rating baseline. All values are in 2015 dollars per Medicare beneficiary. Government spending corresponds to the change in subsidy and rebate payments, including the cost of subsidizing TM (FFS costs). The baseline simulated market share of MA is 27.8%.

uating the designer's objective at the expected equilibrium outcome. Introducing the new scores while keeping quality and prices at their baseline levels delivers about 60% of the surplus and 40% of the profit gains. Allowing quality to change more than doubles consumers' surplus, increases firm profit, and expands MA. Finally, incorporating the equilibrium prices leads to increased premiums and a slight erosion of cost-sharing benefits. These changes create a substantial welfare transfer between consumers and firms, with profits increasing by 62% and surplus dropping by over 56%. Overall, the analysis shows that the welfare effect of quality responses to scores can exceed those of information. It highlights the importance of accounting for equilibrium responses when designing scores.

### 7.3 The full information benchmark

The second column of Table 5 shows the welfare change under a full information counterfactual. As in the theoretical analysis of Section 2, overall welfare is lower than under the best linear substitute design. Analogously, Figure 9a shows that the quality distribution in the regulated market dominates the full information distribution in the first-order stochastic sense. This reveals that the regulator uses scores to induce a higher investment than profit-maximizing for many firms under full information, revealing an aggregate (across quality dimensions) Spencian distortion for MA.

Consumers benefit from the added quality, but so do firms on average. There are two reasons for this. First, as the best linear substitute aggregator is not fully aligned with consumers' preferences, some firms can obtain the same score as plans of higher quality-utility. The misalignment affords them a greater demand than what they would receive under full information for a given investment. Importantly for consumers, this investment is overall larger. Second, as consumers'

Table 6: Welfare decomposition for alternative designs

| | Δ Firm Profits | Δ Consumer Surplus | MA Share | Premium | Benefits |
|---|---|---|---|---|---|
| **Best linear substitue** | | | | | |
| +Δ Information | 191.7 | 113.1 | 56.8% | 21.4 | 71.6 |
| +Δ Quality | 301.4 | 398.4 | 69.3% | 21.2 | 76.1 |
| +Δ Prices | 489.0 | 178.5 | 64.2% | 41.9 | 72.3 |
| **Certification** | | | | | |
| +Δ Information | 83.7 | -9.5 | 44.5% | 23.1 | 68.3 |
| +Δ Quality | 267.1 | 389.2 | 67.5% | 20.9 | 76.1 |
| +Δ Prices | 437.9 | 178.0 | 62.7% | 40.9 | 72.3 |

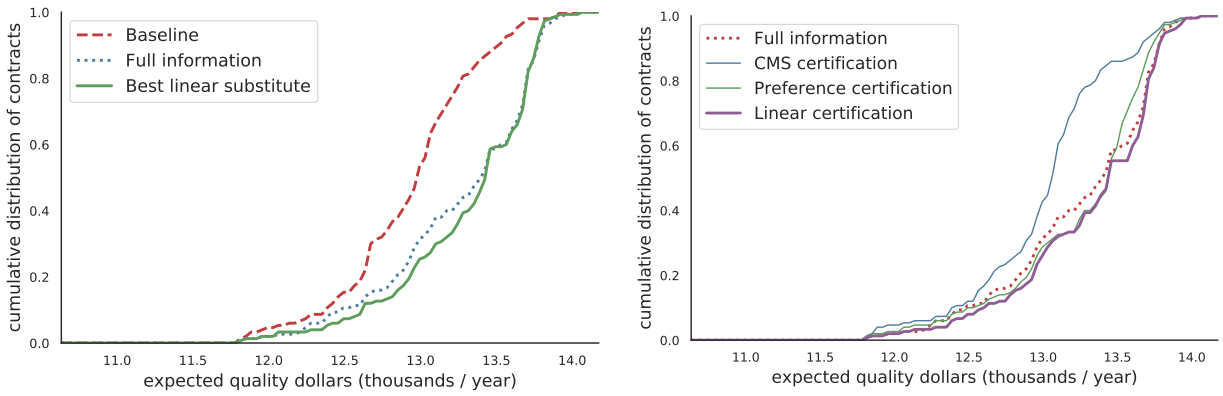
*Notes*: This table displays the welfare change from gradually incorporating the equilibrium outcome of the best linear substitute and the quality certification designs. To avoid confounding the effect of investment risk, this decomposition is done at the expected realized quality given insurers' optimal investments, leading to slightly different numbers than in the preceding table. All values are in 2015 dollars per Medicare beneficiary. Premiums and Benefits are enrollment-weighted averages.

posterior beliefs are interior to the scoring intervals, firms investing near the cutoffs will obtain a greater demand than what they would under full information. In both cases, firms benefit from informational distortions that subtract from consumers' surplus but have a positive net effect.

The finding that a coarse design can outperform full information might have important policy implications. In particular, it shows a contradiction between CMS's stated goals of informing consumers and promoting quality (Medicare Payment Advisory Commission, 2013). The results indicate that when market power over quality exists, providing consumers with detailed quality information might be worse for them and reduce insurers' incentives to invest in quality. The minor welfare differences between the best linear substitute and the full-information benchmark falsely suggest that maximizing the informativeness of scores might be a good heuristic. The following section will show that this is far from true, as the welfare gains from scoring do not increase with informativeness.

### 7.4 Optimal certification

Quality certifications are the most common type of quality disclosure policy. Beyond their simplicity, their prevalence might be due to their ability to affect quality. In particular, certification pools all uncertified low qualities into one score and all high qualities into another. Thus, products below the certification threshold receive the worst possible signal any monotone deterministic




(a) Best linear substitute

(b) Quality certification

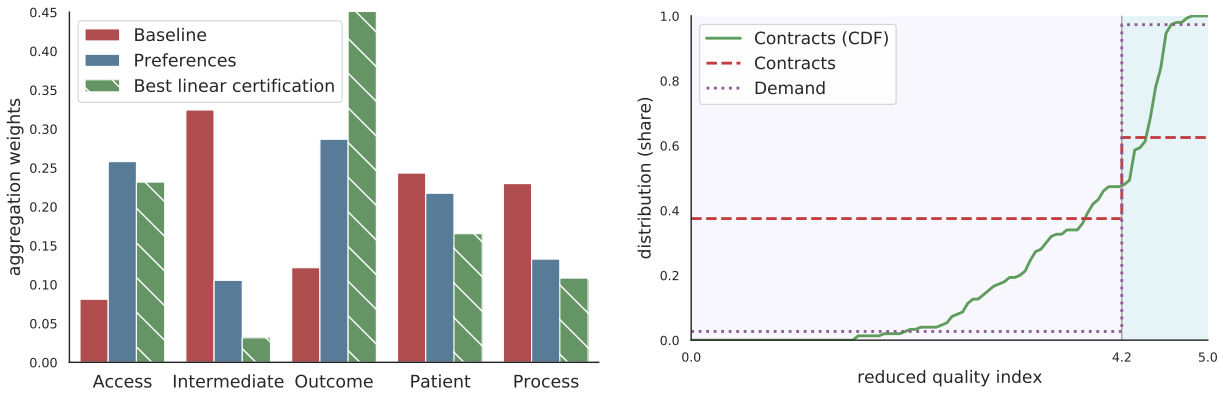
Figure 9: Quality Distribution

*Notes*: These figures display the cumulative distribution of contracts relative to their expected quality. The figure on the left presents the distributions compared in the Best Linear Substitute design, while the figure on the left shows the ones for the certification designs.

score could provide, while those above it obtain the best signal. Intuitively, this will tend to make consumers buy fewer uncertified products, creating incentives for firms to invest in quality. If these incentives are strong enough, few firms will offer uncertified products, and those certified would offer qualities near the certification cutoff to maximize profits. This is precisely the mechanism behind the monopoly regulation example of Section 2, which results in consumers buying a product with exact beliefs about its quality despite obtaining only a dichotomous signal of quality.

To test whether this intuition applies to an empirical context, I solve for the optimal quality certification for MA. I focus on designs that, like the baseline, use a linear function to aggregate quality dimensions and present designs using higher-order aggregators in the appendix. Figure 10 shows the resulting weights and the cutoff placement. As in the previous solution, this design improves the alignment of aggregation weights and consumers' preferences. Also, the certification cutoff is placed in an index that has a similar quality-utility value as the threshold to be top-rated in the best linear substitute. The share of products choosing to receive certification is 61.9%, serving 97.5% of consumers.

The welfare gains from the optimal placement of cutoffs and weights under certification are shown in the third column of table 5. Both firms and consumers prefer simple certification over the more sophisticated status quo. Figure 9b shows that the distribution of quality under certification dominates that of full information, confirming that this design can narrow the Spencian gap. A corollary of these results is that total welfare is non-monotonic in the score's informativeness. However, this is only true because of the endogeneity of quality to the score. Table 6 shows the decomposition of welfare, evaluated the expected quality outcomes. The table shows that replacing




(a) Reduction weights

(b) Quality distribution and cutoffs

Figure 10: Best linear certification design

*Notes*: These figures display the design of the best linear certification scoring system. For further details on the plotted components see the details of figure 7.

the MA Stars with certification while keeping quality and prices fixed would harm consumers. The gains from certification stem exclusively from the quality effect.

The simple structure of quality certification makes it easier to examine the effect of aggregation weights on total welfare. To study this, I solve for the optimal certification while holding CMS's relative weighting scheme fixed and using consumers' preferences to weight categories. That is, the CMS-based certification shows the welfare gains that the market could attain by transitioning to an optimized certification design keeping CMS's priorities over quality as they currently are. Instead, the preference-based certification shows the welfare gains if the designer fully aligned weights with preferences. In this case, the system would guarantee that a higher score implies a higher quality-utility. The computed welfare effects are shown in the last two columns of table 5, and the cutoff placements are in the appendix. Figure 9b contrasts the distribution of quality under the three types of certification.

This exercise indicates that CMS could attain a similar quality distribution and welfare as in the status-quo using quality certification.[51] Under the CMS-based design, 52% of contracts obtain certification, and 91.7% of consumers purchase a certified product. In contrast, perfectly aligning the weights with consumers' preferences results in the same share of certified contracts as under the best linear certification but at a lower threshold. Certified firms reduce their investments by 36% relative to the better certification scheme and serve only 95% of consumers. In general, the results of this work indicate that consumers' willingness to substitute one quality dimension for

[51]This finding is consistent with anecdotal evidence that insurers in MA aim for four stars, suggesting a behavior similar to the one excepted under certification. I learned this from interviewing the director of Star Rating Analytics at one of the largest insurance companies in MA.

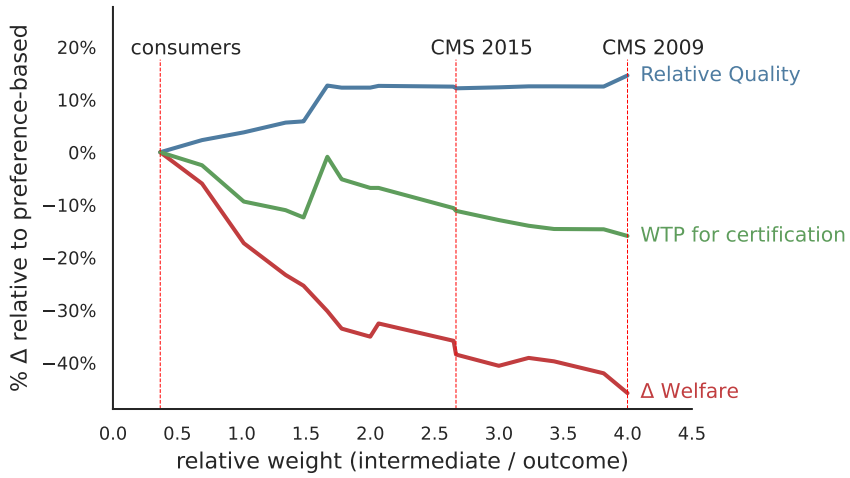



Figure 11: Welfare and certification value under manipulated score

*Notes*: This figure displays how investment, certification value, and welfare change as the score's weight on Intermediate quality relative to Outcome quality increases. The discontinuities are due to discrete changes in the number of firms choosing to certify, leading to abrupt changes in investments.

another serves as an adequate rule-of-thumb replacement for a weighting scheme.

Overall, the results of this section have a clear connection with evidence from other markets and policies in MA. The literature has documented significant supply responses to quality certification in settings such as appliances (Houde, 2018a) and groceries (Barahona et al., 2020). My results suggest that similar incentive schemes can be used to promote quality even in scenarios with imperfect control (i.e., risky investment or effort) and multidimensional quality. It also shows the potential benefit of aligning scoring incentives with consumer preferences in settings where multitasking has been a concern, such as nursing homes (Feng Lu, 2012). Finally, the results highlight the importance of creating large incentives gaps in order to promote quality. Surprisingly, an influential advisory commission recently suggested that Congress remove such incentives – or what they call "cliff effects" – from other quality promoting programs in MA (MedPAC, 2020). This section indicates that cliff effects are crucial to promoting quality and might also improve plan quality information.

### 7.5 Strategic score manipulation

A caveat of the previous analysis is that the designer might value quality beyond its impact on total welfare or subsidy spending. As CMS bears the brunt of the cost if consumers' health deteriorates, the designer and consumers might disagree about the value of reducing long-term health risks. This conjecture could explain why the baseline weighting scheme differs from consumer preferences.

Nevertheless, it is unclear whether there are gains from nudging consumers away from their preferred choices. If the regulator distorts the score to highlight specific products, consumers might respond less to the signal, reducing supply incentives and potentially offsetting the gains from the nudge.

To evaluate this possibility, I solve for different certification equilibria starting from the preference-based design and adjusting the weights of the Intermediate and Outcome categories. I choose the weights to span CMS's designs between 2009 and 2019, keeping the overall contribution of the two categories fixed. The results, shown in Figure 11, indicate that increasing the relative contribution of Intermediate relative to Outcome increases the relative quality investment produced and purchased. However, the change in quality plateaus rapidly at around 12.5%.[52] The reason is that consumers' willingness to pay for certified products deteriorates rapidly as the signal becomes less representative of the information they seek.[53]

Overall, the results suggest that scores are not practical mechanisms for consumer nudging. To justify the observed weighting schemes, the regulator would have to value a small reallocation of quality across categories by hundreds of millions of dollars. As the reallocation implied by CMS's weights is away from Outcomes, it is difficult to assess whether meaningful quality-of-life effects of improving Intermediate quality justify such a distortion.

## 8 Robust scoring design

In this final section, I explore how to design scores when consumers' preferences for quality are unknown. I consider the case in which the designer only knows that preferences are within a particular set denoted $\Gamma$. As appendix theorem 1 shows, this would be the case if consumers did not understand the scoring design changes in the MA market, and instead had exogenous and unchanging beliefs about what a star rating means. In this scenario, a designer lacking alternative sources of elicited preferences and beliefs, would be unable to identify consumers' beliefs about quality but could derive bounds about their preferences for different dimensions. Knowing only the set $\Gamma$, I consider the design objective that maximizes total welfare under the

[52] These percentages are relative to the preference-based certification design. This certification is fast to compute and makes it easier to illustrate the effect of distorting the alignment. Thus, the ratio of Intermediate to Outcome quality of the average chosen product increased by 13% relative to the same ratio under the preference-based certification equilibrium.

[53] The level of certified quality (not shown) also decreases because of dampened demand incentives, dropping from \$16 to \$15.5 dollars per month in ex-post utility. Uncertified quality remains stable at \$13.4.

worst-case preferences:[54]

$$\max_{\psi\in\Psi}\min_{\gamma\in\Gamma}\int\left(\underbrace{CS(\psi,\boldsymbol{q})}_{\substack{\text{Consumer}\\\text{surplus}}}+\rho^{F}\underbrace{\sum_{f}V_{f}(\psi,\boldsymbol{q})-I(\boldsymbol{x}_{f}^{*}(\psi),\mu_{f})}_{\substack{\text{Insurer}\\\text{profit}}}-\underbrace{\rho^{G}Gov(\psi,\boldsymbol{q})}_{\substack{\text{Government}\\\text{spending}}}\right)dF(\boldsymbol{q}|\boldsymbol{x}^{*}(\psi)) \tag{13}$$

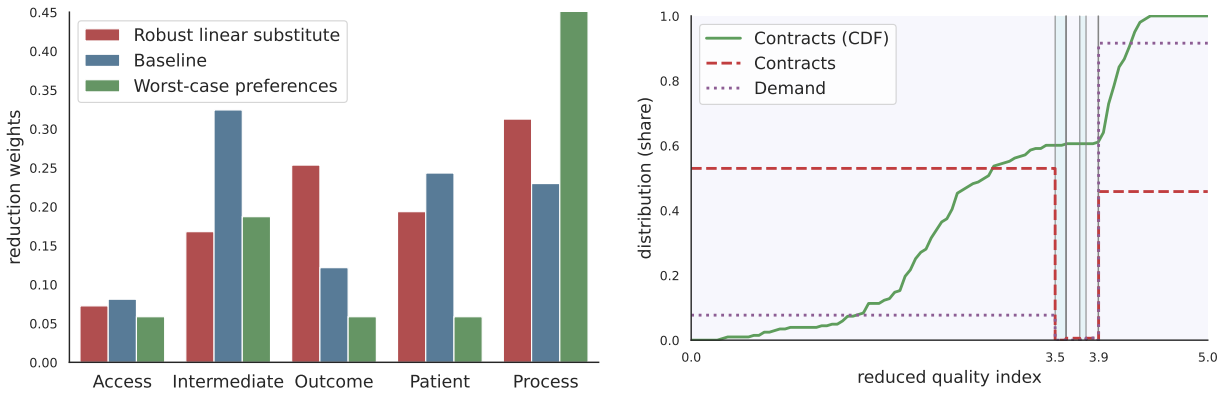
As consumers' beliefs are unknown and independent of the design, the designer's toolkit is limited to using the nine pre-existing scores. Her objective is to use these scores to maximize welfare under the worst-case scenario of preferences. Importantly, this objective would remain the same if consumers had heterogeneous preferences for quality. This cautious approach matches the decisions of an imperfectly informed regulator that risks significant political or legal losses from implementing a new design that worsens outcomes. Appendix Proposition 5 shows that the interior minimization is equivalent to a linear equilibrium constraint, which facilitates solving this problem.

I solve for the linear reduction and associated cutoffs that optimize the robust design problem.[55] As in the main analysis, I consider two solutions, one using all nine scores and another using only one and five stars to create a certification scheme. The computed design for the *robust linear substitute* is shown in Figure 12 and the certification design is shown in the associated appendix section. Intuitively, given that consumers' preferences are adversarial, the design attempts to promote uniform quality production. The non-uniform bounds on preferences and firms' cost heterogeneity skew the optimal weight placement, resulting in a higher relevance for the Process, Patient Experience, and Outcome categories than for Access and Intermediate Outcomes. The cutoff structure of this design is peculiar, resulting in a "padded" certification scheme. As in standard certification, the system assigns one star to all low qualities and five stars to all high ones. The padding of this design is that it allocates five intermediate stars in the gap between the bottom and top scores. In expectation, no firm lands in this narrow gap. However, it does increase firm profit by reducing their investment risk when attempting to reach the top score.

The fundamental mechanisms of the robust design are the same as with informed choice. Consumers penalize lower-rated products leading to an increase in investment and certification.

[54] Preferences $\gamma$ are relevant only for consumer surplus as, conditional on the identified fixed valuation for scores, the demand is independent of consumers' preferences for quality.

[55] In order to discipline the worst-case preferences, I further restrict $\Gamma$ such that $\gamma$ are between half the lowest estimated value and twice the highest among all quality dimensions. This means that the highest quality product can be worth anywhere between \$4133 and \$44984 per year in premiums. Otherwise, the worst-case scenario often derives zero utility from the quality dimension with the highest investment, which is unreasonably harsh.




(a) Reduction weights

(b) Quality distribution and cutoffs

Figure 12: Robust certification design

*Notes*: These figures display the reduction weights, worst-case preferences, and cutoff placement of the robust linear substitute design. The figure on the left displays the reduction weights in the baseline and the new design. The worst-case preferences for this problem are also illustrated. The figure on the right shows the cutoff placements and the resulting distribution of contracts.

In equilibrium, top-rated products in both robust designs deliver a quality that is very similar in value to those of five-star plans in the baseline. However, a larger share of plans obtain this score and serve about 91% of consumers. Thus, the robust designs eliminate a range of intermediate qualities that distort firms' incentives without significantly benefiting consumers. The vertical nature of quality and the limited cost heterogeneity makes it welfare-enhancing to trim these intermediate qualities.

The welfare effect of the robust linear substitute and the certification designs are displayed in Table 7. The estimates indicate that both improve the worst-case scenario welfare. As the more flexible design results in a certification scheme, the welfare difference between the two solutions is minimal. In both cases, as consumers' preferences are unknown and adversarial, the consumer surplus gain is mechanically small. However, the results still indicate an expansion of MA and a reduction in public spending.

The robust design objective can also be used to evaluate the worst-case preference scenario for the designs derived under the assumption of informed choice. The last two columns of Table 7 display these effects for the best linear substitute and linear certification design. Both have a positive welfare impact, indicating that even if they were designed with wrong conjectures about consumer preferences for quality, they would still lead to overall improvements. This indicates that redesigning the system to jolt investment and competition is almost always beneficial, even if the relative importance of quality investments is erroneously measured. This finding, of course, is driven by quality being a vertical attribute of products: all else fixed, more quality is always better

Table 7: Worst-case welfare changes in redesigned system

| | **Robust** | | **Previous Designs** | | |
|---|---|---|---|---|---|
| | Substitute | Certification | Substitute | Certification | CMS-based |
| Δ Consumer Surplus | 5.6 | 12.7 | -322.9 | -249.4 | 72.1 |
| Δ Firm Profits | 288.6 | 277.0 | 428.7 | 415.1 | 173.0 |
| Δ Gov. Spending | -154.6 | -146.1 | -163.2 | -175.5 | -107.9 |
| Δ Total Welfare | 294.3 | 289.7 | 105.7 | 165.7 | 245.1 |
| MA share | 50.7% | 49.6% | 60.0% | 57.0% | 42.0% |

*Notes*: This table displays the welfare effect of redesigning the scoring system, relative to the MA Star Rating baseline under the worst-case consumer preferences for quality. The baseline for welfare comparison also evaluates the worst-case preferences given the Star Ratings. The Previous Designs columns evaluate the worst-case for the designs derived using the informed choice assumption. The CMS-based columns presents the quality certification design that uses CMS's weighting scheme. All values are in 2015 dollars per Medicare beneficiary. Government spending corresponds to the change in subsidy and rebate payments, including the cost of subsidizing TM (FFS costs).

in every dimension.

Finally, the last column of Table 7 shows the worst-case scenario for the CMS-based certification design. This design uses CMS's weighting scheme to aggregate quality dimensions, one of the main differences between the robust and main certification designs. The results indicate that the CMS weighting scheme is particularly well suited to address the robust problem, although the cutoff placements could be improved. In particular, CMS's weights are nearly optimal within the class of linear reductions and result in a better worst-case scenario than both the main linear substitute and certification designs. This observation suggests that the CMS weighting decision might be driven by an abundance of caution about misrepresenting consumers' preferences.

Overall, this exercise complements the work of the previous sections in three ways. First, it helps to disentangle the mechanisms by which the score affects the market. In particular, in the previous sections, the designs coordinated consumers by changing the assignment of scores to products and their beliefs about the quality represented by those scores. In this exercise, the second channel is eliminated, showing that scoring design can be effective even if consumers are unaware of design changes. Second, it highlights the importance of creating transparent and well-communicated scoring systems. The gap between the results of this section and the previous – consumers' understanding of the design – appears fixable by an informational regulator. Finally, it provides an alternative solution for the cautious regulator (or reader) unnerved by the assumption of informed choice. Moreover, it allows me to compute a worst-case scenario for the previously proposed designs.

# 9 Conclusion

I study the problem of designing a scoring system for firms with market power over quality. Using data from Medicare Advantage in 2009-2015, I show that scores shift demand across products and alter insurers' quality investments. Leveraging individual-level choices and variation in the scoring design, I estimate a structural model of demand and supply responses to scoring. I specify the problem of a welfare-maximizing designer and use the model estimates to evaluate alternative designs and find local optima. The analysis presents three novel findings.

First, I show that the optimal disclosure policy for MA involves coarsening quality information. Total welfare is neither increasing in how informative a scoring system is nor is it maximized at full information. The central mechanisms for this behavior are that quality responds to scores and is underprovided by firms under full information. Scores can marshal demand to offset firms' market power over quality, increasing total welfare. I propose an alternative design that increases welfare substantially, with half of the improvement stemming from better quality information and the remainder from the endogenous quality responses. This finding highlights the importance of considering the effect of information policies on the endogenous supply of quality and evaluating how these might be coordinated or overlap with other efforts to alter the market. This result also provides empirical support to the growing theory on scoring design (Rodina and Farragut, 2016; Ball, 2019; Hopenhayn and Saeedi, 2019; Boleslavsky and Kim, 2018; Zapechelnyuk, 2020).

Second, I find that a well-designed quality certification can vastly improve welfare and tightly approximate the effect of more sophisticated scores. In MA, I find that certification achieves 94% of the welfare gains of an optimized nine-scores system. The results indicate that cliff-effects in firms' incentives are fundamental in promoting quality, contradicting some recent policy recommendations (MedPAC, 2020). This finding also highlights the contradiction between CMS's effort to inform consumers, promote quality, and improve overall welfare in the market. Instead, my results show that scoring designs that reveal very little information can result in more informed purchases of higher quality. Finally, these results also provide evidence on why some certification schemes have been exceptionally effective despite their simplicity (Barahona et al., 2020).

Third, I find that skewing the score's information away from what consumers care about quickly erodes its value, and thus its ability to alter the market's outcome. This finding has important implications for scoring designs in general, particularly for CMS's practice of seeking design feedback from the industry. The results suggest that CMS should instead elicit consumers' preferences for quality when designing its quality aggregation scheme. A combination of theoretical and empirical results in this paper also highlights the importance of clear communication and transparency in scoring design. They play an essential role in eliciting consumers' preferences and

in the score's effectiveness as both an information and regulatory policy.

My results point the way to several possible extensions. Extending this analysis to incorporate market dynamics would be helpful for policy design. Also, incorporating informational frictions to the designer's problem, and more importantly, data manipulation as in Ball (2019) would help extend these tools to markets where that has been an issue, such as nursing homes (Silver-greenberg and Gebeloff, 2021). Finally, I assume that the quality domain and dimensions are fixed. Allowing investments to exceed observed qualities or the designer to choose from new dimensions might also prove valuable.

## References


Abaluck, J., Caceres Bravo, M., Hull, P., and Starc, A. (2021). Mortality Effects and Choice Across Private Health Insurance Plans. *The Quarterly Journal of Economics*, 136(3):1557–1610.

Ackerberg, D. a. (2009). A new use of importance sampling to reduce computational burden in simulation estimation. *Quantitative Marketing and Economics*, 7(4):343–376.

Aizawa, N. and Kim, Y. S. (2018). Advertising and risk selection in health insurance markets. *American Economic Review*, 108(3):828–867.

Albano, G. L. and Lizzeri, A. (2001). Strategic certification and provision of quality. *International Economic Review*, 42(1):267–283.

Allende, C., Gallego, F., and Neilson, C. (2019). Approximating the Equilibrium Effects of Informed School Choice. *Working Paper*, (1100623).

Angrist, J. D. and Guryan, J. (2008). Does teacher testing raise teacher quality? Evidence from state certification requirements. *Economics of Education Review*, 27(5):483–503.

Anscombe, F. J. and Aumann, R. J. (1963). A Definition of Subjective Probability. *The Annals of Mathematical Statistics*, 34(1):199–205.

Araya, S., Elberg, A., Noton, C., and Schwartz, D. (2018). Identifying Food Labeling Effects on Consumer Behavior. *SSRN Electronic Journal*.

Atal, J., Cuesta, J. I., and Saethre, M. (2021). Quality Regulation and Competition: Evidence from Pharmaceutical Markets.

Baker, A., Larcker, D. F., and Wang, C. C. Y. (2021). How Much Should We Trust Staggered Difference-In-Differences Estimates? *SSRN Electronic Journal*, (March).

Ball, I. (2019). Scoring Strategic Agents. (January):1–63.

Barahona, N., Otero, C., Otero, S., and Kim, J. (2020). Equilibrium Effects of Food Labeling Policies.

Barrios, J. M. (2017). Occupational Licensing and Accountant Quality: Evidence from Linkedin. *SSRN Electronic Journal*.

Berry, S., Eizenberg, A., and Waldfogel, J. (2016). Optimal product variety in radio markets. *RAND Journal of Economics*, 47(3):463–497.

Berry, S. and Haile, P. (2020). Nonparametric Identification of Differentiated Products Demand Using Micro Data. *National Bureau of Economic Research*.

Berry, S., Levinsohn, J., and Pakes, A. (1995). Automobile Prices in Market Equilibrium. *Econometrica*, 63(4):841.

Berry, S. T. (1994). Estimating Discrete-Choice Models of Product Differentiation. *The RAND Journal of Economics*, 25(2):242.

Berry, S. T. and Waldfogel, J. (2001). Do mergers increase product variety? Evidence from radio broadcasting. *Quarterly Journal of Economics*, 116(3):1009–1025.

Boleslavsky, R. and Kim, K. (2018). Bayesian Persuasion and Moral Hazard. *SSRN Electronic Journal*, (March).

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

Brown, J., Duggan, M., Kuziemko, I., and Woolston, W. (2014). How does risk selection respond to risk adjustment? New evidence from the Medicare Advantage Program. *American Economic Review*, 104(10):3335–3364.

Brown, Z. Y. (2018). An Empirical Model of Price Transparency and Markups in Health Care. *Working Paper*, (August).

Charbi, A. (2020). The fault in our stars! Quality Reporting, Bonus Payments and Welfare in Medicare Advantage*.

Chen, Y. (2018). User-Generated Physician Ratings-Evidence from Yelp.

Chernew, M., Gowrisankaran, G., and Scanlon, D. P. (2008). Learning and the value of information: Evidence from health plan report cards. *Journal of Econometrics*, 144(1):156–174.

Chou, S. Y., Deily, M. E., Li, S., and Lu, Y. (2014). Competition and the impact of online hospital report cards. *Journal of Health Economics*, 34(1):42–58.

Clay, K., Severnini, E., and Sun, X. (2021). Does LEED Certification Save Energy? Evidence from Federal Buildings.

CMS (2016). Quality Strategy. Technical report.

Cooper, Z., Gibbons, S., Jones, S., and Mcguire, A. (2011). Does hospital competition save lives? Evidence from the English NHS patient choice reforms. *Economic Journal*, 121(554):228–260.

Crawford, G. S., Shcherbakov, O., and Shum, M. (2019). Quality overprovision in cable television markets †. *American Economic Review*, 109(3):956–995.

Crawford, G. S. and Shum, M. (2005). Uncertainty and learning in pharmaceutical demand. *Econometrica*, 73(4):1137–1173.

Curto, V., Einav, L., Finkelstein, A., Levin, J., and Bhattacharya, J. (2019). Health care spending and utilization in public and private medicare. *American Economic Journal: Applied Economics*, 11(2):302–332.

Curto, V., Einav, L., Levin, J., and Bhattacharya, J. (2021a). Can health insurance competition work? Evidence from medicare advantage. *Journal of Political Economy*, 129(2):570–606.

Curto, V., Einav, L., Levin, J., and Bhattacharya, J. (2021b). Can health insurance competition work? Evidence from medicare advantage.

Cutler, D. M., Huckman, R. S., and Kolstad, J. T. (2010). Input constraints and the efficiency of entry: Lessons from cardiac surgery. *American Economic Journal: Economic Policy*, 2(1):51–76.

Dafny, L. (2010). Are Health Insurance Markets Competitive? *American Economic Review*, 100(3):1399–1431.

Dafny, L. and Dranove, D. (2008). Do report cards tell consumers anything they don't already know? The case of Medicare HMOs. *RAND Journal of Economics*, 39(3):790–821.

Darden, M. and McCarthy, I. M. (2015). The star treatment: Estimating the impact of star ratings on medicare advantage enrollments. *Journal of Human Resources*, 50(4):980–1008.

Dranove, D. and Dafny, L. (2008). Do Report Cards Tell Consumers Anything They Don ' T Already. *RAND Journal of Economics*, 39(3):790–821.

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

Dranove, D. and Sfekas, A. (2008). Start spreading the news: A structural estimate of the effects of New York hospital report cards. *Journal of Health Economics*, 27(5):1201–1207.

Dworczak, P. and Martini, G. (2019). The simple economics of optimal persuasion. *Journal of Political Economy*, 127(5):1993–2048.

Elfenbein, D. W., Fisman, R., and McManus, B. (2015). Market structure, reputation, and the value of quality certification. *American Economic Journal: Microeconomics*, 7(4):83–108.

Fan, Y. (2013). Ownership consolidation and product characteristics: A study of the US daily newspaper market. *American Economic Review*, 103(5):1598–1628.

Fan, Y. and Yang, C. (2020). Competition, product proliferation, and welfare: A study of the US smartphone market. *American Economic Journal: Microeconomics*, 12(2):99–134.

Farronato, C., Fradkin, A., Larsen, B., and Brynjolfsson, E. (2020). Consumer Protection in an Online World: An Analysis of Occupational Licensing. *National Bureau of Economic Research Working Paper Series*, No. 26601.

Feng Lu, S. (2012). Multitasking, Information Disclosure, and Product Quality: Evidence from Nursing Homes. *Journal of Economics and Management Strategy*, 21(3):673–705.

Forbes, S. J., Lederman, M., and Tombe, T. (2015). Quality disclosure programs and internal organizational practices: Evidence from airline flight delays. *American Economic Journal: Microeconomics*, 7(2):1–26.

Fox, J. T., il Kim, K., Ryan, S. P., and Bajari, P. (2011). A simple estimator for the distribution of random coefficients. *Quantitative Economics*, 2(3):381–418.

Frank, R. G. and McGuire, T. G. (2019). Market Concentration and Potential Competition in Medicare Advantage. *Issue brief (Commonwealth Fund)*, 2019(February):1–8.

Gandhi, A., Froeb, L., Tschantz, S., and Werden, G. J. (2008). Post-merger product repositioning. *Journal of Industrial Economics*, 56(1):49–67.

Gaynor, M., Moreno-Serra, R., and Propper, C. (2013). Death by market power: Reform, competition, and patient outcomes in the national health service. *American Economic Journal: Economic Policy*, 5(4):134–166.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*.

Goolsbee, A. and Petrin, A. (2004). The consumer gains from direct broadcast satellites and the competition with cable TV. *Econometrica*, 72(2):351–381.

Handel, B. R. (2013). Adverse selection and inertia in health insurance markets: When nudging hurts. *American Economic Review*, 103(7):2643–2682.

Harbaugh, R. and Rasmusen, E. (2018). Coarse grades: Informing the public by withholding information. *American Economic Journal: Microeconomics*, 10(1):210–235.

Ho, K. and Handel, B. (2021). INDUSTRIAL ORGANIZATION OF HEALTH CARE MARKETS. *NBER Working Paper*.

Ho, K. and Lee, R. S. (2017). Insurer Competition in Health Care Markets. *Econometrica*, 85(2):379–417.

Holmstrom, B. and Milgrom, P. (1991). Multitask Principal–Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design. *The Journal of Law, Economics, and Organization*, 7(special_issue):24–52.

Hopenhayn, H. and Saeedi, M. (2019). Optimal Ratings and Market Outcomes.

Horowitz, J. L. and Markatou, M. (1996). Semiparametric Estimation of Regression Models for Panel Data. *Review of Economic Studies*, 63(1):145–168.

Houde, S. (2018a). Bunching with the Stars: How Firms Respond to Environmental Certification. *SSRN Electronic Journal*, (July).

Houde, S. (2018b). The Incidence of Coarse Certification: Evidence from the Energy Star Program. *SSRN Electronic Journal*.

Hui, X., Saeedi, M., Spagnolo, G., and Tadelis, S. (2018). Certification, Reputation, and Entry: An Empirical Analysis. *NBER Working Paper*, 24916.

Izbicki, R. and Lee, A. B. (2017). Converting high-dimensional regression to high-dimensional conditional density estimation. *Electronic Journal of Statistics*, 11(2):2800–2831.

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

Jin, Y. and Vasserman, S. (2019). Buying Data from Consumers.

Kamenica, E. (2019). Bayesian Persuasion and Information Design. *Annual Review of Economics*, 11(1):249–272.

Kamenica, E. and Gentzkow, M. (2011). Bayesian persuasion. *American Economic Review*, 101(6):2590–2615.

Kleiner, M. and Soltas, E. (2019). A Welfare Analysis of Occupational Licensing in U.S. States. *National Bureau of Economic Research Working Paper Series*, (1122374).

Kolstad, J. T. (2013). Information and quality when motivation is intrinsic: Evidence from surgeon report cards. *American Economic Review*, 103(7):2875–2910.

Krasnokutskaya, E. (2011). Identification and estimation of auction models with unobserved heterogeneity. *Review of Economic Studies*, 78(1):293–327.

Larsen, B. (2014). Occupational Licensing and Quality: Distributional and Heterogeneous Effects in the Teaching Profession. *SSRN Electronic Journal*, (0645960):1–52.

Larsen, B., Ju, Z., Kapor, A., and Yu, C. (2020). THE EFFECT OF OCCUPATIONAL LICENSING STRINGENCY ON THE TEACHER QUALITY DISTRIBUTION. *National Bureau of Economic Research Working Paper Series*.

Lustig, J. (2010). Measuring welfare losses from adverse selection and imperfect competition in privatized medicare. *Manuscript. Boston University Department of . . .* , pages 1–49.

McGuire, T. G., Newhouse, J. P., and Sinaiko, A. D. (2011). An economic history of Medicare Part C. *Milbank Quarterly*, 89(2):289–332.

McManus, B. (2007). Nonlinear pricing in an oligopoly market: The case of specialty coffee. *RAND Journal of Economics*, 38(2):512–532.

Medicare Payment Advisory Commission (2013). The Medicare Advantage Program: Status Report. *Report to Congress: Medicare Payment Policy*, pages 287–306.

MedPAC (2020). The Medicare Advantage Program: Status Report. *Report to Congress: Medicare Payment Policy*, pages 287–306.

Miller, K. S., Petrin, A., Town, R., and Chernew, M. (2019). Optimal Managed Competition Subsidies. *National Bureau of Economic Research Working Paper Series*, No. 25616.

Mizala, A. and Urquiola, M. (2013). School markets: The impact of information approximating schools' effectiveness. *Journal of Development Economics*, 103(1):313–335.

Mussa, M. and Rosen, S. (1978). Monopoly and product quality. *Journal of Economic Theory*, 18(2):301–317.

Neal, D. and Schanzenbach, D. W. (2010). Left behind by design: Proficiency counts and test-based accountability. *Review of Economics and Statistics*, 92(2):263–283.

Newhouse, J. P. and McGuire, T. G. (2014). How successful is medicare advantage? *Milbank Quarterly*, 92(2):351–394.

Nosal, K. (2011). Estimating Switching Costs for Medicare Advantage Plans.

Nosko, C. (2014). Competition and quality choice in the cpu market. *Manuscript, Harvard University*, (November):1–54.

Reid, R. O., Deb, P., Howell, B. L., and Shrank, W. H. (2013). Plan Star Ratings and Enrollment. *Journal of the American Medical Association*, 309(3):267–274.

Rodina, D. and Farragut, J. (2016). Inducing Effort Through Grades. *Working Paper*.

Ronnen, U. (1991). Minimum Quality Standards , Fixed Costs , and Competition Author ( s ): Uri Ronnen Published by : Wiley on behalf of RAND Corporation Stable URL : http://www.jstor.org/stable/2600984 Minimum quality standards , fixed costs , and competition. *The RAND Journal of Economics*, 22(4):490–504.

Ryan, C. (2020). How does Insurance Competition Affect Medical Consumption?

Schennach, S. M. (2016). Recent Advances in the Measurement Error Literature.

Schmalensee, R. (1979). Market structure, durability, and quality: a selective survey. *Economic Inquiry*, XVII.

Silver-greenberg, J. and Gebeloff, R. (2021). Maggots, rape and yet five stars: How u.s. ratings of nursing homes mislead the public. The New York Times https://www.nytimes.com/2021/03/13/business/nursing-homes-ratings-medicare-covid.html. Accessed: 06/26/2021.

Small, K. A. and Rosen, H. S. (1981). Applied Welfare Economics with Discrete Choice Models. *Econometrica*, 49(1):105.

So, J. (2019). Adverse Selection, Product Variety, and Welfare.

Spence, A. M. (1975). Monopoly , Quality , and Regulation. *The Bell Journal Of Economics*, 6(2):417–429.

Sweeting, A. (2009). The strategic timing incentives of commercial radio stations: An empirical analysis using multiple equilibria. *RAND Journal of Economics*, 40(4):710–742.

Town, R. and Liu, S. (2003). The Welfare Impact of Medicare HMOs. *The RAND Journal of Economics*, 34(4):719.

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

Werner, R. M., Norton, E. C., Konetzka, R. T., and Polsky, D. (2012). Do consumers respond to publicly reported quality information? Evidence from nursing homes. *Journal of Health Economics*, 31(1):50–61.

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