Quality Disclosure and Regulation: Scoring Design in Medicare Advantage [*]

Benjamin Vatter[†]

September 19, 2022

Abstract

Policymakers and market intermediaries often use quality scores to alleviate asymmetric information about product quality. Scores affect the demand for quality and, in equilibrium, its supply. Equilibrium effects break the rule whereby more information is always better, and the optimal design of scores must account for them. In the context of Medicare Advantage, I find that consumers' information is limited, and quality is inefficiently low. A simple design alleviates these issues and increases consumer surplus by 2.4 monthly premiums. More than half of the gains stem from scores' effect on quality rather than information. Scores can outperform full-information outcomes by regulating inefficient oligopolistic quality provision, and a binary certification of quality attains 94% of this welfare. Scores are informative even when coarse; firms' incentives are to produce quality at the scoring threshold, which consumers know. The primary design challenge of scores is to dictate thresholds and thus regulate quality.

Keywords: disclosure, quality regulation, information design, equilibrium effects, welfare, competition

JEL Codes: L15, L11, I11, I18, D82, D83

[*] First version: October 2021. 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, Samuel Goldberg, Eilidh Geddes, Piotr Dworczak, Hugo Hopenhayn, Philip Haile, and seminar participants at Northwestern University and several other institutions for their valuable comments and suggestions. This work benefited from generous funding from the Robert Eisner Graduate Fellowship. Any errors are my own.

[†] Stanford University. Email: bvatter@stanford.edu

I Introduction

Quality scores are ubiquitous. From car emissions to school performance, regulators and certifying agencies rely on scores for disclosure. Scores help consumers choose when information is scarce and, by doing so, also alter firms' incentives to invest in quality. While a growing theoretical literature provides valuable guidelines for designing disclosure policies, their welfare-optimal design depends on empirical fundamentals such as consumers' willingness to pay for quality, the degree of quality competition, and firms' ability to adjust to disclosure. The wrong design can exacerbate information frictions, distort firms' incentives, and even harm consumers.[1] Due to such concerns, much of the empirical literature on disclosure evaluates scores' ambiguous impact. In this paper, I bridge theory and empirics by studying optimal design in a real-world setting, estimating its primitives, examining the gains from alternate designs, and quantifying the relative importance of different design choices.

The effect of scores on the supply of quality breaks the rule that more information is always better for consumers (Blackwell, 1953). Coarser information can benefit consumers by regulating inefficiencies in quality provision caused, for example, by externalities in R&D, production subsidies, or limited competition. This paper focuses on the latter, specifically, on *Spencian* distortions (Spence, 1975) caused by firms' inability to capture surplus created by marginal quality increments from inframarginal consumers. When these distortions are not competed away, scores can coordinate demand to penalize inefficient firms and, simultaneously, reveal to consumers whether products are of efficient quality. Hence, in equilibrium, scores can lead to efficiency in quality and information.[2]

I apply these ideas to study the Medicare Advantage (MA) Star Rating health insurance scores. This policy assigns plans a score between 1 and 5 stars, in half-star increments, according to their performance along five quality dimensions.[3] The MA setting provides a valuable laboratory for studying disclosure design: The rules mapping quality measurements to scores—i.e., the scoring design—vary annually, the regulator's quality measurement data are readily available for all plans, and there are no competing sources of quality scores for consumers. Moreover, firms are incentivized to compete on quality because their revenue is risk-adjusted, and premiums are highly regulated and subsidized. It is also an important setting in its own right: There are 65 million Medicare beneficiaries, and quality impacts

[1] For an example of harm in the nursing home industry, see Silver-greenberg and Gebeloff (2021).

[2] See the example in Section II for an illustration of this idea.

[3] These are preventive care, access to care, medical quality, chronic condition management, and patient experience. I detail the design in Section III.

mortality (Abaluck *et al.*, 2021) and entails billions in public spending (CMS, 2016).

I document three fundamental observations about scoring design in MA. In Section IV, I show that, first, consumers have monotonically increasing preferences for scores: New enrollees are 20% more likely to choose a 5-star than a 2-star plan, all else equal. Second, consumers' preferences correlate with changes to the mapping between quality and scores: The preference for 5 over 2-star plans depends on which qualities are awarded 5 instead of 2 stars. Third, firms respond to design changes by adjusting quality rapidly and proportionally to the scoring incentives. These observations, and the variation that underlies them, reveal consumers' preferences for scores and firms' quality production costs. Through the lens of a model, these can predict the impact of counterfactual scoring policies.

In Section V, I develop a model of quality investment, plan pricing, and enrollment. The model captures four key frictions: two informational and two dimensions of moral hazard. First, consumers' information about quality is limited to scores, so they cannot distinguish between the qualities of equally rated plans. Second, unless the scoring design aggregates quality dimensions precisely as consumers' preferences, consumers cannot tell whether a higher-rated plan has a preferred aggregate quality over a lower-rated one. This *across-scores* distortion, and the previous *within-scores* distortion, can lead consumers to suboptimal enrollment decisions.

The first dimension of principal-agent moral hazard affects the provision of aggregate quality. The regulator would like heterogeneous insurers to provide a variety of efficient quality-price combinations for consumers who differ in willingness to pay (WTP) for quality. Firms' incentive, however, is to maximize profits. The regulator must choose a common contract for all firms: a scoring policy. Scores determine the set of potential market qualities since firms' incentive is to attain scores at the lowest cost, precisely at the scoring threshold (Kolotilin and Zapechelnyuk, 2019). More scores increase the potential for valuable variety but also inefficient deviations in production. The second type of moral hazard affects the relative allocation of aggregate quality investments across quality dimensions. Since consumers cannot ascertain by which combination of qualities a plan obtained its score, firms ignore consumers' preferences over quality dimensions. Instead, they consider only their costs and relative impact on scores, which conflicts with the regulator's preferences for efficiency and creates a multitasking moral hazard problem (Holmstrom and Milgrom, 1991).

In Section VI, I show that the primitives underlying the informational and moral hazard frictions are identified from variation in MA's design, enrollment, and quality. Consumers' WTP for scores is identified from the trade-off between premiums and scores in enrollment. Their preferences and beliefs about plan quality are identified from the correlation of WTP and changes to the scoring design. The same variation identifies firms' investment costs because it changes the relative gains of investing in different quality dimensions.

Model estimates reveal that quality is inefficiently provided, and consumers' information is severely limited. A marginal increase in the median contract's aggregate quality increases each enrollee's surplus by $3,051 more than it costs to produce. However, of the five quality dimensions, two are overprovided. Efficiency is heterogeneous, with some contracts having excess or lacking quality in every dimension. On the consumers' side, information frictions reduce their surplus by approximately three monthly premiums. Across-score frictions account for 95% of this loss since 22.4% of plans are *misclassified* from consumers' perspective; there is a lower scoring alternative of higher aggregate quality.

I use the model estimates and a novel methodology to find an alternative, constrained optimal design for MA.[4] The new system is a simple discretization of plans' weighted average qualities into five scoring levels (four fewer than the Star Ratings), which relies on three features. First, all medium-to-low qualities are pooled at the bottom score. Pooling decreases consumers' expectations of plan quality and induces a demand penalty for underprovision, which lessens the Spencian distortion. Second, more scoring levels are assigned to higher qualities, which balances product variety and efficiency and reduces within-score informational frictions. Third, the averaging weights are optimized to better align with consumers' preferences while accounting for heterogeneity in firms' costs, which lowers across-score frictions and multitasking moral hazard. This final feature is the most important; a binary certification with optimal weights attains 94% of the constrained optimum's welfare. Thus the granularity of scores—their most visible and discussed design choice—is the least welfare relevant once we account for equilibrium supply responses.

The alternative increases consumer surplus by $146.5 per Medicare beneficiary per year and total welfare by $669.2. Design changes affect consumers' information, product prices, and quality. Changing information alone reduces the mean squared error of consumers' beliefs about quality by 91.7%, which increases surplus by $44.8. The change in information reveals firms' vertical differentiation, which, keeping quality constant, leads to 37.5% higher markups on the average chosen plan and a surplus transfer of $57.1 from consumers to firms. Allowing quality to adjust increases aggregate quality by 4.3%, surplus by $183.2, and total welfare by 185.9%; quality regulation is the main driver of the scores' welfare gain.

Accounting for equilibrium effects drastically changes the optimal scoring design. With

[4] The constraint is to the space to which the Star Ratings belong. This is the class of all designs that deterministically assign a higher quality to weakly greater scores, using finitely many scoring levels.

exogenous quality, welfare under full information is about 58% higher than under the new coarse scoring system. Equilibrium quality effects overturn the dominance of full information, and welfare under the new scores is 7% higher due to the regulation of Spencian distortions. Distortions decrease with the number of competitors, and the gains from coarsening information vanish once markets have about five competing firms, holding vertical differentiation constant. In equilibrium, only 9.9% of consumers would be better off having full information about product quality than under the alternative scoring system.

These results could be interpreted as a recommendation for the Centers for Medicare and Medicaid Services (CMS). But given the substantial gains provided by the alternative design, one might wonder why CMS's policies have systematically differed. One explanation is that CMS's objective includes factors beyond those considered here, such as the cost of subsidized care. I consider the possibility of such private regulatory preferences in Section VII.C.4, but find that CMS would have to value minor improvements in quality by exceedingly large values. A more compelling alternative is given in Section VIII, in which I study the problem of robust scoring design. In the robust case, the designer is uncertain about consumers' quality preferences and cannot affect their beliefs. This is a worst-case scenario for disclosure, and accordingly, it seeks to maximize the worst-case welfare. CMS's design is nearly optimal in such a setting, which suggests CMS might be highly cautious about misrepresenting consumer preferences. Given Medicare's delicate political and social role, this finding is, perhaps, reasonable. Nevertheless, I show that the optimal robust design can still do better by leveraging the same economic forces described above to induce higher quality and better matching between consumers and products.

Related Literature. This exercise in empirical scoring design bridges a gap between the theoretical literature on the subject and the empirical literature that measures disclosures' impact.[5] To my knowledge, few papers have explored this gap. Dai *et al.* (2018) study the optimal aggregation of subjective consumer restaurant reviews, while one of the counterfactual exercises in Barahona *et al.* (2022) explores optimal certification for ready-to-eat cereals. This paper extends these ideas to the broader agenda on information design with moral hazard (Boleslavsky and Kim, 2018) by examining optimal granularity, aggregation, and the trade-off between quality and informational regulation.

[5] On the theoretical side, these include Albano and Lizzeri (2001); Glazer and McGuire (2006); Harbaugh and Rasmusen (2018); Ball (2020); Hopenhayn and Saeedi (2019); and Zapechelnyuk (2020). The latter is particularly relevant, as it considers market power and moral hazard in scoring. On the empirical side, the literature includes Jin and Sorensen (2006); Elfenbein *et al.* (2015); Araya *et al.* (2018); Alé-Chilet and Moshary (2022); Reynaert and Sallee (2021); and Charbi (2020), who measures the welfare value of the MA Star Ratings. See Dranove and Jin (2010) for a review of earlier work on quality disclosure and Kamenica (2019) for closely related work on theoretical information design.

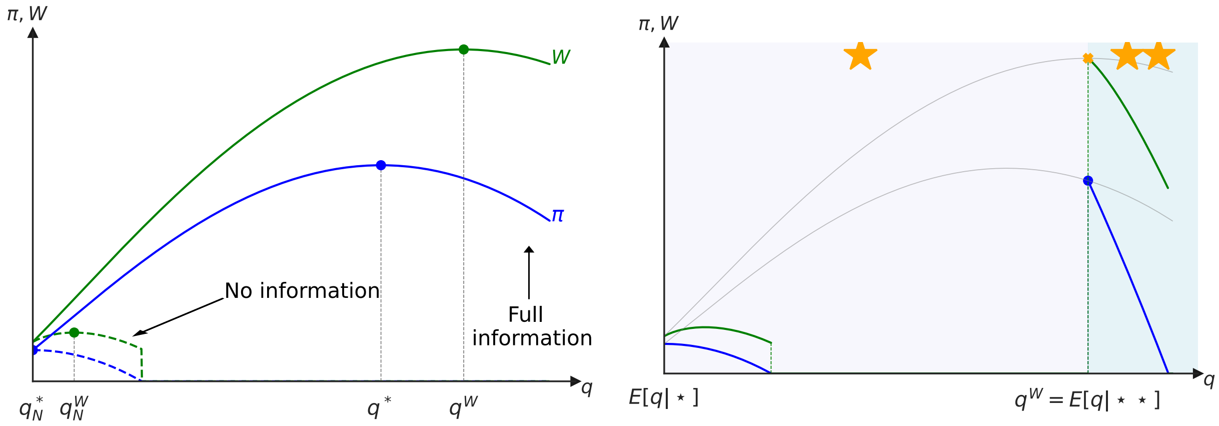
My results show that MA scores can act as effective quality regulation, which contributes to research on the supply effects of centralized mandatory disclosure (Jin and Leslie, 2003; Houde, 2018; Allende *et al.*, 2019; Barahona *et al.*, 2022) and the empirical study of quality regulation broadly (Angrist and Guryan, 2008; Kleiner and Soltas, 2019; Larsen *et al.*, 2020; Atal *et al.*, 2022). My examination of the regulation of imperfect competition among insurers expands on the literature on quality provision in healthcare markets (Cutler *et al.*, 2010; Cooper *et al.*, 2011; Gaynor *et al.*, 2013; Kolstad, 2013; Fleitas, 2020) and competition among insurers (Ho and Lee, 2017; Ho and Handel, 2021). In particular, I quantify the effects of moral hazard in quality provision among insurers competing for the demand of incompletely informed consumers. I prove that consumers' priors and quality preferences can be identified from design variation, which contributes to the study of choice under incomplete information (Abaluck and Gruber, 2011; Chernew *et al.*, 2008; Handel and Kolstad, 2015).

This paper connects research on the industrial organization of Medicare Advantage (Town and Liu, 2003; Lustig, 2009; Aizawa and Kim, 2018; Curto *et al.*, 2021; Nosal, 2011; So, 2019; Charbi, 2020; Miller *et al.*, 2022; Ryan, 2020; Fioretti and Wang, 2019; Decarolis *et al.*, 2020a) to the literature on insurance market design (Handel *et al.*, 2015; Marone and Sabety, 2022; Decarolis *et al.*, 2020b). I study the role of purely informational policies, whose implementation often focuses on statistical issues and maximizing informativeness. I provide evidence that their design must consider equilibrium supply effects and that doing so can drastically change the optimal solution. Closely related, Miller *et al.* (2022) study optimal subsidies and competition over coverage generosity in Medicare Advantage. This paper is complementary and extends the policy analysis to disclosure and competition over quality.

II Disclosure as Quality Regulation

I begin by describing the economic intuition underlying scores' ability to regulate quality while informing consumers.[6] Consider a single-product monopolist selling an indivisible good. The monopolist chooses price and quality and pays a production cost that increases in its quality q. A regulator observes q and discloses a public score (or signal) $\psi(q)$. Consumers cannot observe q before purchase but know the scoring rule $\psi(\cdot)$ and the resulting score. Using this information and knowing that quality is costly to produce, consumers form rational expectations about q and make purchasing decisions. The regulator seeks to maximize welfare by committing to a public scoring rule before the monopolist's quality is chosen.

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



(a) Informational extrema

(b) Regulated

Figure 1: Quality certification under monopolistic provision

Notes: These figures illustrate how certification changes a monopolist's investment incentives. Figure (a) presents profit and welfare curves for the full and no-information scenarios. The latter vanishes when profits are negative, leading to a market exit. Figure (b) presents how profit and welfare change when consumers are only informed whether quality exceeded q^w. Shaded areas illustrate the distinct scores.

There are two informational extrema attainable by the scores. On the one hand, a constant score reveals no information to consumers, which renders demand inelastic to quality and thus eliminates any incentive for the monopolist to invest in it. On the other hand, a fully informative score allows the monopolist to exert market power over quality (Crawford *et al.*, 2019), leading to potentially inefficient investment (Spence, 1975). Intuitively, when evaluating a marginal increase in quality under full information, the monopolist considers its effect on the marginal consumer. The regulator, instead, accounts for the surplus created by the increase for both marginal and inframarginal consumers. Hence, the monopolist's quality choice will likely be inefficient, even when consumers are fully informed of product quality.[7] Figure 1a illustrates these two extrema and the resulting inefficiencies.

The regulator can address these inefficiencies by using coarse scores. For example, Figure 1b illustrates the outcome of a scoring rule that only certifies whether the quality is at least efficient ($q \geq q^W$). This policy disrupts the firm's profit curve because, on both sides of the certification cutoff, demand is inelastic to quality—as in the no-information extremum—but at different levels. To the left, consumers are guaranteed a ceiling on the good's quality. Knowing that quality is costly to produce and that the insurer lacks any incentive to provide quality interior to the interval, they expect $q = 0$. To the right, consumers are guaranteed that quality is at least q^W and, by the same logic as before, expect $q = q^W$. Therefore,

[7] As noted by Spence (1975), this inefficient quality production can lead to over- or underprovision. Efficient output is also feasible under particular demand forms, such as linear.

if the monopolist's full-information profits at $q = 0$ are lower than at $q = q^W$, it will invest efficiently when regulated. Consumers' expectations would be accurate, so this design eliminates market power over quality and informational distortions.

This illustration reveals that scores can improve on full-information outcomes by acting as quality-regulation policies. The regulatory power of scores stems from their ability to marshal demand and coordinate consumers to offset the monopolist's market power. This mechanism extends well beyond this simple example. Scores can shift demand even if consumers have biased priors, have to choose among multiple products, or face other sources of uncertainty. The solution often differs from a dichotomous certification since detailed scores are more informative 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. In the following sections, I develop an empirical methodology to recover these components and systematically translate these inputs into optimal scoring designs.

III Institutional Details and Data

III.A 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 Centers 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. Over the last decade, MA has become increasingly popular and covered 48% of the 65 million Medicare-eligible beneficiaries in 2022.[8]

MA markets are highly concentrated and regulated. In 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 various plans that differ in coverage generosity (e.g., coinsurance, deductibles) and access to clinical quality. CMS regulates the financial characteristics of plans, including minimum

[8] 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 more information on risk adjustment and residual selection, see Brown *et al.* (2014) and So (2019).

requirements on coverage generosity and limits on premiums relative to coverage (Curto *et al.*, 2021). CMS also subsidizes enrollees' premiums, resulting in zero premiums for nearly half of all MA plans.[9]

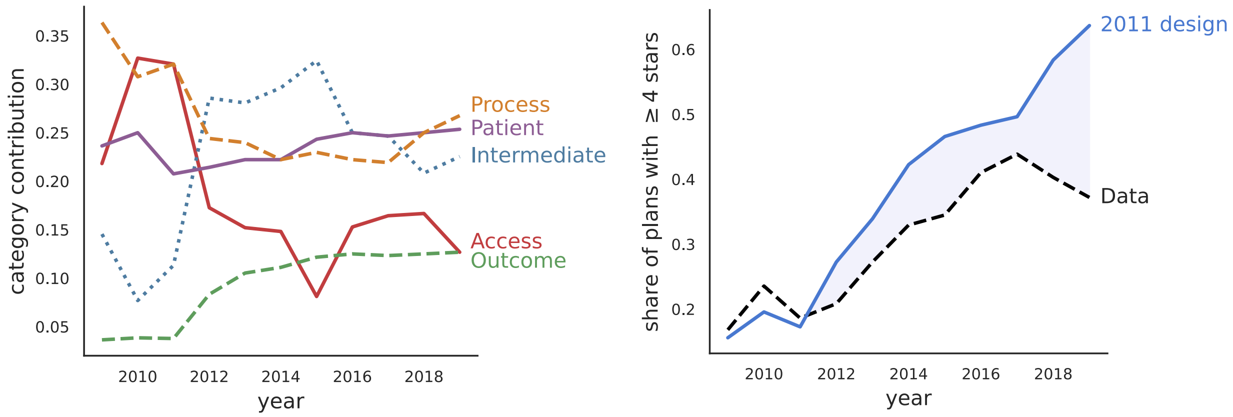
In contrast to financial characteristics, differences in plan quality are less regulated and harder for consumers to ascertain. These differences are due to variations across plans in the size and makeup of provider networks, disease management protocols, and processes for approving costly medical procedures, among other factors. Since insurers can offer the same network arrangement and services under different cost-sharing and premium combinations, CMS measures quality at the *contract* level. A contract is a group of plans from the same insurer that (according to CMS) share the same quality. The median contract has only two plans, with 70% of its enrollment in one of them, and the median consumer observes only one of a contract's plans. Therefore, in many cases, the distinction between plan and contract is irrelevant. However, having price variation conditional on quality and year will prove helpful for estimation purposes. Throughout, I refer to products in MA as "plans" and use the term "contract" only when relevant for clarity or exposition.

Information regarding the quality of plans is rarely available to consumers when choosing coverage. To assist consumers, CMS created the Star Ratings scoring system, which displays a coarse summary of each plan's quality next to the enrollment button in Medicare's unified shopping platform.[10] To compute these scores, CMS first collects information on over 60 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 1 to 5 to each plan-measure, ascending in quality. Next, CMS chooses a weight for each category and computes a weighted average of all measure-level scores for each plan. Overall, denoting w_k the weight of each category $k \in \mathcal{K}$ and $\mathcal{L}_k$ the measurements included in the category, the score of plan j is given by

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

[9] MA consumers pay a Part B premium regardless of their choice of TM or MA. For further details regarding the MA market regulation, see Online Appendix I.

[10] See Online Appendix Figure 1 for a view of the online platform. For a description of earlier quality scores in MA, see Dafny and Dranove (2008).



(a) Design variation

(b) Variation in score assignment

Figure 2: Scoring design variation and simulated assignment under constant design

Notes: Figure (a) shows the evolution of category contributions to the design. For each category, the contribution is the product of its number of measurements (e.g., Process measures include breast cancer screening and kidney disease monitoring) and its weight, divided by the total weight among all measurements. Figure (b) shows the change in the scoring assignment if CMS had kept its 2011 scoring design, keeping quality as measured in the data. The shaded area highlights the gap across the resulting assignments. Adjustment factors are preserved as measured in the corresponding year.

Where $\text{Round}_{.5}(\cdot)$ rounds a number to its nearest half and q_{kj} is the quality of plan j in measure k. The adjustment factor, ω_j, captures minor bonuses due to past performance.[11]

CMS frequently changed the weights and number of measures in each category, introducing substantial variation in the Star Rating design. In 2012, CMS moved from uniform weights to a 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 yearly as CMS experimented with measures. Overall, each category's contribution to the scores has varied significantly, as shown in Figure 2a. As I detail in Online Appendix I, consumers likely observed this variation since the composition of categories was visible on the Medicare website and enrollment platform.

Design variation significantly impacted score assignment. Figure 2b shows that if CMS had kept its 2011 scoring design, 60% of plans in 2019 would have received 4 or more stars, while the actual number was 40%. The difference is due primarily to a decrease in the importance of the Access category and an increase in that of both the Outcome and Intermediate Outcome categories. Thus, in 2011 a high-scoring plan afforded consumers excellent access to physicians and a median-quality network of hospitals. In 2019, the roles of hospital quality and access to physicians were reversed. The figures also show an improvement in overall

[11] See the supplementary material for full construction details.

quality as the share of top-rated plans increases under a constant design.

Finally, CMS provides dynamic incentives. Starting in 2012, plan subsidies and scoring adjustment factors depend on past quality performance.[12] However, this paper aims to understand the short-run mechanisms, effects, and design of a purely informational quality disclosure policy. Thus, I incorporate dynamic features as they appear in the data and treat them as sources of heterogeneity. I exclude pecuniary incentives from the designer's toolkit to avoid confusing gains from information design with those from direct transfers.

III.B Data

This paper combines three data sources; the first is plan-market-level data from 2009 to 2019. Each year, CMS publishes a data compendium that contains information on every MA plan in each county. I use it to construct a panel of plans, their market-level enrollment, subsidies, prices, rebates, premiums, plan benefits, and cost-sharing. The data also provide the total number of Medicare-eligible beneficiaries in each county and information regarding the dual Medicare-Medicaid eligible population. I use these data to adjust the sample by removing dual eligibles and plans specifically designed for that population.[13]

The second data source is the Medicare Current Beneficiary Survey (MCBS). This nationally representative rotating panel tracks around 15,000 Medicare beneficiaries annually for up to 4 years. I obtain data that covers 2009 to 2015 and provides information on individual demographics, well-being, income, location, and enrollment choices.[14] I restrict the data to non-dual beneficiaries within the continental US and who have geographic information, which leaves 46,833 beneficiary-years. The panel also provides sampling weights to compare the survey's demographics with the national population. However, the data do not include all counties. This limits my primary welfare analyses to about 22 million individuals or approximately one-third of the overall Medicare population.

Finally, the third data source pertains to plans' quality and the scoring rules. Each year, CMS publishes the data used to compute the star ratings, including quality measurements,

[12] MA also reward plans that achieve 5 stars by allowing consumers to switch into them after the open enrollment period ends. Since few 5-star plans exist, 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. A few insurers exploited this practice to combine contracts to manipulate their scores for a year.

[13] Similar restrictions have been used by Aizawa and Kim (2018), Miller *et al.* (2022) and Curto *et al.* (2021). I present descriptive statistics in the Online Appendix Table 5. Construction details are available as supplementary material.

[14] Excluding 2014, because it was never released to the public due to implementation difficulties.

assigned scores, and cutoffs. The data, however, do not explain changes to underlying measurement scales, weights, or variable definitions. To address this, I completed the data by reviewing a decade of public communications by CMS aimed at insurers. I recover year-to-year changes to the scoring design and perfectly replicate the public scoring assignment.

IV Descriptive Evidence on Market Responses to Scoring

Scores' regulatory power stems from their effect on demand and, consequently, firms' investment incentives. However, whether the Star Ratings can influence demand and supply is an empirical question. For example, scores might be irrelevant if they summarize information consumers already know or fail to affect firms if their production technology is immutable. This section provides empirical evidence that MA Star Ratings affected both sides of the market and have, therefore, the potential to be a powerful regulatory tool.

IV.A Enrollment

Medicare beneficiaries' response to scores when making enrollment decisions has been thoroughly documented in previous work (Dafny and Dranove, 2008; Reid *et al.*, 2013; Darden and McCarthy, 2015).[15] This effect of scores is easy to observe in the individual enrollment panel by examining whether, all else equal, consumers prefer higher- to lower-scoring plans. While focusing on decisions made by new enrollees among MA plans, I regress an indicator of their choices on the score assigned to each plan, controlling for all other observable (to consumers) attributes of plans, including premiums, coverage, and additional benefits:[16]

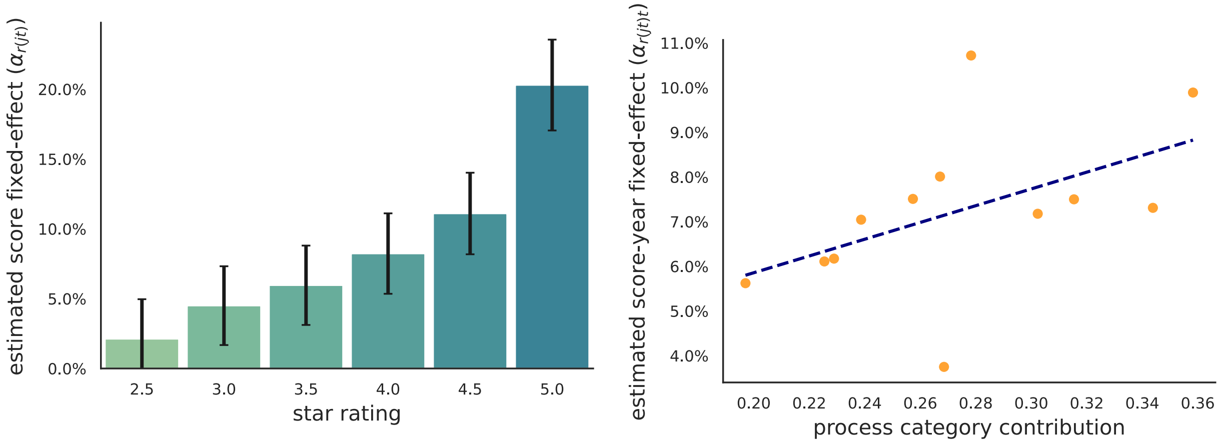
$$y_{ijt} = \alpha_{r(jt)} + \boldsymbol{x}_{jt}\boldsymbol{\lambda} + \mu_{m(i)} + \xi_t + \epsilon_{ijt} \quad . \tag{2}$$

Above, y_{ijt} indicates that consumer i chose plan j in year t; $\alpha_{r(jt)}$ is a fixed effect for the score of plan j in year t; and $\boldsymbol{x}_{jt}$ denotes plan characteristics, ξ_t a year fixed effect, and $\mu_{m(i)}$ a market fixed effect. Figure 3a displays the estimates of $\alpha_{r(jt)}$, the coefficient of interest.

The results demonstrate that all else equal, consumers prefer higher-scoring plans. A 5-star plan is approximately 20% more likely to be chosen by a new enrollee than an equivalent 2-star plan (the normalized category), conditional on the consumer's choosing an MA plan.

[15] The first two articles use aggregate enrollment data, while the third uses cross-sectional individual-level data. Here, I rely on individual-level panel data to select consumers potentially unaffected by inertia.

[16] As TM is not scored and inertia in plan choice is well documented for this market (Nosal, 2011), I focus on choices made by enrollees new to MA, conditional on choosing an MA plan. Locked-in consumers might fail to switch from deteriorating plans even if they value quality and scores. These switching costs are likely partly caused by the hassle of changing physicians and treatment interruptions (Drake *et al.*, 2022).



(a) Conditional on design

(b) Across designs relative to Process category

Figure 3: Estimated effect of scores on plan choice probability, conditional on MA

Notes: Figure (a) displays the estimated $\alpha_{r(jt)}$ from equation (2). Error bars indicate 95% confidence intervals. The normalized category is 2 stars. Figure (b) shows a binned scatter plot of estimated score-year fixed effects against the contribution of the Process category, controlling for the score level.

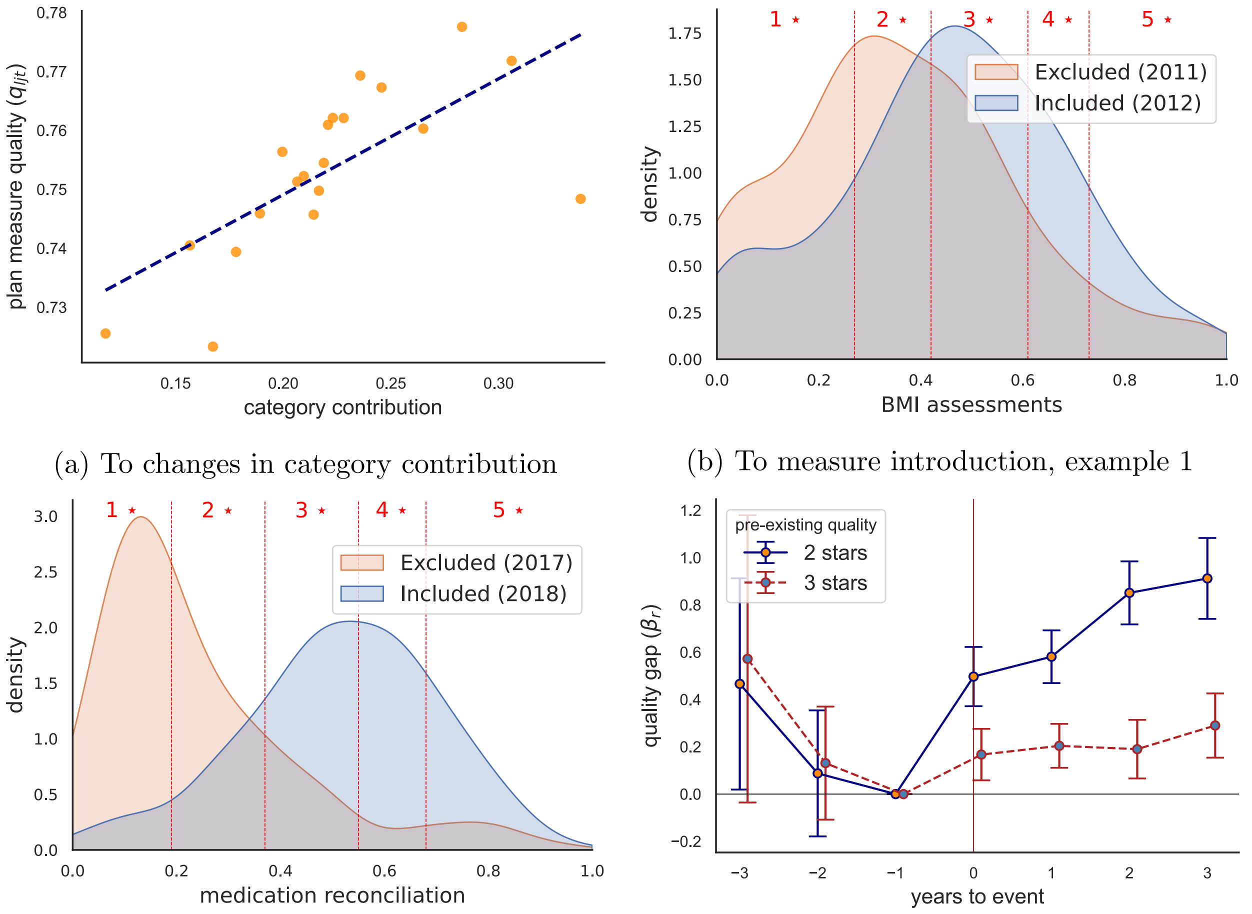
Scores' effect on demand is monotonic—as expected if consumers understand that scores signal quality. However, the informational content of scores varies with the design. For example, a plan excelling in Medical Outcome quality while being average for everything else would receive 3 stars in 2011 but 4 in 2012.

To evaluate responses to design changes, I modify equation (2) to estimate score-year fixed effects, $\alpha_{r(jt)t}$, dropping the year fixed effects. Figure 3b shows that these new estimates vary substantially and correlate with category contributions. This analysis controls for the star level, which suggests that consumers' interpretation of scores varies with the design.[17] Together, these results suggest that scores shift consumers across products and do so differently depending on their design.

IV.B Quality

The first suggestive evidence that plan quality responds to the incentives produced by scores is their correlation with category contributions. This is shown in Figure 4a, which illustrates how the quality of plans in any given measure positively relates to its category's contribution to the overall score. The variation is due to plans' quality changes rather than their market composition since I control for heterogeneity across plans and categories (interacted).

[17] This figure, however, does not imply that consumers value having more information about Process quality, as the changes across categories are correlated. The structural model accounts for these correlated changes and their effect on the informational content of scores.



(a) To changes in category contribution

(b) To measure introduction, example 1

(c) To measure introduction, example 2

(d) To measure introduction, event study

Figure 4: Plan quality response to design variation

Notes: Figure (a) plots binned plan-measure quality relative to its category's contribution, conditional on a contract-category fixed effect. Figures (b) and (c) display the distribution of quality for two measures introduced to the design during the study period. Vertical lines mark the measure-level scoring bins in the introduction year. The horizontal axis marks the frequency with which a plan performs the quality process. Figure (d) shows estimates of equation (3), relative to the 4-star category and the year before introduction. Bars mark 95% confidence intervals.

However, this correlation might stem from policy adjustments to exogenous quality shifts.

To explore the causal link, I examine the response of quality measures to their introduction to the design. This variation is a small subset of the factors that cause changes in category contribution, but it has several advantages. First, for many measures, CMS evaluated their quality before and after the design change. Second, these changes were announced to insurers without anticipation.[18] Finally, because the scoring rule converts quality

[18] Changes were announced a year before measurement, allowing insurers to respond in time.

measurements to measure-level scores, the change produced clear and heterogeneous incentives across firms. For example, Figures 4b and 4c show the distribution of two measures introduced in 2012 and 2018, respectively. In the first example, plans with a quality of 0.1 in 2011 faced the risk of getting a 1 added to their overall weighted average quality score if they failed to improve by 2012. In contrast, those with preexisting quality above 0.7 had no such incentive, as their measure-level score in 2012 would be 5 regardless. The same logic applies to the second example and seven other such events in the data.

I apply this logic to compare the evolution of quality across dimensions, plans, and time using a triple-difference regression. I assume preexisting heterogeneity in excluded quality measures was independent of the unanticipated change in design and that firms were on similar trends across the thresholds. Therefore, plans of high preexisting quality follow the trend that low preexisting quality ones would have followed if not for the change in design.

$$\underbrace{q_{ljt}}_{\substack{\text{normalized quality}\\ \text{measurement}}} = \underbrace{\sum_r \beta_r T_{lt} \mathbb{1}\left\{G_{lj} = r\right\}}_{\substack{\text{introduced measure}\\ \text{x preexisting quality score}}} + \underbrace{\gamma_{lj} + \mu_{lt} + \xi_{jt}}_{\text{pairwise fixed effects}} + \epsilon_{ljt} \tag{3}$$

Above, T_{lt} indicates if measure l is included in the scoring design at year t, and G_{lj} equals the measure-level score of each plan-measure using the design of the year of introduction but the quality of the preceding year.[19] To compare quality metrics, I standardize them using their means and standard deviations across all years. To avoid conflating the effects of bounded quality domains, I drop plans in the first and last preexisting quality groups and normalize the coefficient of interest (β_r) for the fourth group to zero.[20]

Three differences are involved in the analysis. First is a comparison within plan-measure, controlled for by the fixed effect γ_{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 fixed effect μ_{lt}, makes the comparison across groups. In this case, β_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 ξ_{jt}. This accounts for the evolution of overall quality in each plan and the trend in MA. Thus the analysis compares the quality change

[19] For example, in Figure 4b, I classify a plan of measure quality 0.5 in the third group.

[20] The domain of most quality measures is bounded (e.g., the share of enrollees receiving a treatment). Therefore, low-quality plans can only improve, and high-quality ones can only worsen, and a failure to account for this would inflate the measurements of this analysis. See Online Appendix I for robustness.

in plan-measures, accounting for general trends in quality in each dimension and plan. The coefficient of interest is identified from variation in quality within measures across time and its differential evolution across quality groups.

Figure 4d plots the evolution of β_r over time, and Online Appendix I presents the underlying regression table and robustness to common methodological concerns. The figure shows that before the design change, plans evolved similarly across the spectrum of preexisting quality. However, once incentives changed, plans of low preexisting quality improved substantially. Within a year, plans in the second group closed on average 29.6% of the gap between the 2-star and 5-star thresholds. Plans in the third group also improve, closing 18.7% of their gap with 5-stars. In both cases, firms respond immediately, and further improvements are small and not statistically significant.

This exercise reveals that consumers respond to scores by changing enrollment decisions and firms by adjusting their quality. These adjustments happen quickly and vary depending on the stakes firms have in responding.[21] The following section presents a model of the market that rationalizes these scoring effects and allows me to leverage MA's extensive variation further. In particular, I use variation in product characteristics and enrollment 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 enrollment costs. The evolution of scoring designs changes firms' incentives to invest in quality and reveals their costs. The same variation reveals consumers' valuation for scores, thus pinning their quality preferences and beliefs.

V Model

I model insurance provision and enrollment as the Perfect Bayesian equilibrium of a series of repeated static interactions between consumers and insurers. At the beginning of each year, the regulator discloses a national quality scoring rule. Insurers then simultaneously choose investments that stochastically determine plans' qualities. They then set plan prices, 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. Next, I present the game's last three stages in reverse order, omitting the regulator's choice stage since I do not impose optimality conditions on the observed scores. I discuss the model's central assumptions at the end.

[21] CMS did not select measures randomly. Therefore, the results do not speak to the effect of scoring a generic quality dimension. The variation needed to measure these effects exists in the data, as scoring rules change yearly, yet disentangling them from the overall data variation requires a more structured approach.

V.A Demand

Building on Town and Liu (2003), 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^{total}_{jmt}, cost-sharing benefits level b_{jmt}, additional plan attributes $\boldsymbol{a}_{jmt}$ (e.g., bundled dental insurance), and a score of r_{jt}. Consumers maximize a Von Neumann-Morgenstern expected utility, evaluating subjective beliefs over quality. The expected indirect utility of consumer i from plan j is

$$u_{ijmt} = \underbrace{\alpha_i p^{\text{total}}_{jmt}}_{\text{premium}} + \underbrace{\beta_i b_{jmt}}_{\text{benefits}} + \underbrace{\mathcal{E}[v(\boldsymbol{q})|r_{jt}, \psi_t]}_{\text{quality}} + \underbrace{\boldsymbol{\lambda}^{a\prime}\boldsymbol{a}_{jmt}}_{\substack{\text{plan}\\\text{attributes}}} + \underbrace{\boldsymbol{\lambda}^{l\prime}\boldsymbol{l}_{ijt}}_{\substack{\text{lock-in}\\\text{indicators}}} + \underbrace{\xi_{jmt}}_{\substack{\text{unobserved}\\\text{preference}}} + \underbrace{\varepsilon_{ijmt}}_{\sim T1EV} \tag{4}$$

Consumers have heterogeneous preferences for premiums and benefits (α_i, β_i). Following Curto *et al.* (2021), both variables are dollar-valued, with the latter being the expected dollars saved from insurance, according to CMS. This metric summarizes all cost-sharing attributes of the plan, such as copayments and coinsurance rates, and is shown to consumers on the enrollment platform.[22] Consumers 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 (ψ_t). Enrollment choices are also affected by the plan's bundled services $(\boldsymbol{\lambda}^a)$ and by previous relationships with firms and products in the market $(\boldsymbol{\lambda}^l)$. Following Handel (2013), this last factor captures MA enrollment inertia (Nosal, 2011) as a direct utility impact. Finally, consumers have unobserved preferences for plans (ξ_{jmt}) and independent type-1 extreme value idiosyncratic preferences (ε_{ijmt}), as in Aizawa and Kim (2018) and Miller *et al.* (2022).

Consumers can also opt for TM coverage. Since most MA enrollees choose plans including prescription drug coverage, I assume they would also bundle TM with a Part D prescription drug plan. I denote by b_0 TM's standard insurance benefits and p^D_{0mt} the price of the market's most popular Part D plan. Consumers' heterogeneous preferences for the outside option are captured by $\boldsymbol{\lambda}^d$. The outside option's utility is thus

$$u_{i0mt} = \underbrace{\alpha_i p^D_{0mt}}_{\text{premium}} + \underbrace{\beta_i b_0}_{\text{benefits}} + \underbrace{\boldsymbol{\lambda}^{d\prime}\boldsymbol{dem}_{it}}_{\substack{\text{consumer}\\\text{demographics}}} + \varepsilon_{i0mt} \tag{5}$$

Given this model, the expected demand for product j in market m in year t is the sum of

[22] Consumers in MA and TM must pay a fixed Part B premium. Since this premium is common across all options, I normalize it in this exposition. Premiums are shown separately for items such as drug plans and insurance, even if bundled. Benefits are shown to consumers as expected out-of-pocket payments, while CMS evaluates these as insurer payments for regulatory purposes. I use the regulatory value such that a plan's benefit level grows with generosity.

the probabilities with which each consumer chooses the product.

$$D_{jmt} = \sum_{i \in \mathcal{I}_m} s_{ijmt} = \sum_{i \in \mathcal{I}_m} \underbrace{\frac{\exp(\delta_{ijmt})}{\exp(\delta_{i0mt}) + \sum_{j' \in \mathcal{J}_{mt}} \exp(\delta_{ij'mt})}}_{\text{individual choice probability}}, \tag{6}$$

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

V.B Supply

V.B.1 Insurers' pricing problem: Each year t, at the third stage of the game, insurance firm f observes the vectors of realized qualities $\boldsymbol{q}_t$ and scores $\psi_t(\boldsymbol{q}_t) = \boldsymbol{r}_t$. Given this information, the firm chooses prices to maximize its total profits given by[23]

$$V_{fmt}(\boldsymbol{q}_t, \psi) = \max_{\boldsymbol{p}_{fmt}} \sum_{j \in \mathcal{J}_{fmt}} \underbrace{D_{jmt}(\boldsymbol{p}_{mt}, \boldsymbol{r}_t)}_{\text{demand}} RA_{jmt} \left(\underbrace{p_{jmt} + R(p_{jmt}, \boldsymbol{z}_{jt})}_{\text{marginal revenue}} - \underbrace{C(\boldsymbol{q}_{jt}, \boldsymbol{a}_{jmt}, \boldsymbol{\theta}^c)}_{\text{marginal cost}} \right) \tag{7}$$

Each plan's demand is multiplied by the risk-adjustment factor RA_{jmt}, which CMS determines for the plan before this stage. The plan's marginal revenue is the sum of its price (p_{jmt}) and additional revenue from prescription coverage and subsidies ($R(\cdot)$). The latter depends on the plan's price and attributes ($\boldsymbol{z}_{jt}$), which include its counties of service and the share of benefits the plan finances with subsidies. I present the formula for this function and how prices map to premiums and benefits in Online Appendix II. $C(\cdot)$ contains the cost of covering each enrollee's standard Medicare benefits, prescription drugs, non-Medicare extra benefits (e.g., dental insurance), and management. 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 ($\boldsymbol{\theta}^c$), the only unknowns of this stage of the model.

Premium and benefit regulations introduce a kink in the demand and revenue of a firm as a function of prices. If the firm sets prices above the kink—known as 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.

[23] In MA, this price is called a "bid." I avoid this terminology to prevent confusion with auctions.

V.B.2 Insurers' investment problem: In the game's second stage, each firm observes the regulator's scoring rule ψ_t and chooses an investment level x_{ckt} for each of its contracts $c \in \mathcal{C}_{ft}$ and category of quality k.[24] 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' choices maximize their total expected profits:

$$\pi_{ft}(\psi_t) = \max_{\boldsymbol{x}_{ft}} \sum_m \underbrace{\int \mathbb{E}_{mt}[V_{fmt}(\boldsymbol{q}_f, \boldsymbol{q}_{-f}, \psi_t)] dF(\boldsymbol{q}_f|\boldsymbol{x}_{ft})}_{\text{expected insurance profit}} - \underbrace{I(\boldsymbol{x}_{ft}, \boldsymbol{\mu}_{ft})}_{\text{investment cost}} \tag{8}$$

Profits equal expected insurance profits ($V_{fmt}(\cdot)$) in each market m minus the cost of quality investments ($I(\cdot)$). Costs are known functions up to 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 $F(\boldsymbol{q}_f|\boldsymbol{x}_{ft})$.[25] In practice, insurers form networks and write contracts that attempt to achieve certain targets but might fall short of or exceed their intended goals. This investment risk distribution captures this idea and is an unknown function to be estimated.

Second, firms are uncertain about their rivals' investment costs and, therefore, their choices at this stage. Since rival investments affect the firm's profits only insofar as 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, similar to the assumption made by Sweeting (2009).

V.C Discussion

The model makes two simplifications that might affect the scoring design analysis. First, consumers have homogeneous preferences over quality dimensions ($v(\boldsymbol{q})$), which makes aggregate quality a vertical attribute (Mussa and Rosen, 1978). It also reduces the computational cost of solving the scoring design problem, which is a stochastic optimization over a non-

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

[25] I present evidence of insurers' imperfect quality control in Online Appendix I.

smooth functional space.[26] Nevertheless, this simplification is unlikely to impact the central question of this paper meaningfully because it does not preclude heterogeneity in willingness to pay for aggregate quality. In Online Appendix II, I show that this heterogeneity is sufficient to generate over- and underprovision of quality and thus capture fundamental inefficiencies in quality provision. In Section VIII, I show that 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 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 III and IV supports this claim. Moreover, the largest insurers in MA entered decades ago and have 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.[27] The data do 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 in the next. 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 design due to the switching costs associated with ongoing treatment or illness.

VI Identification and Estimation

VI.A Demand

I estimate the demand model using the two-step approach of Goolsbee and Petrin (2004). The first step uses individual-level enrollment decisions to recover preference heterogeneity and aggregate market shares to estimate mean population preferences. Splitting the premium and benefit parameters in equation (4) into their mean (α, β) and variation $(\tilde{\alpha}_i, \tilde{\beta}_i)$, the method aggregates mean preferences with all common components of a plan's utility—including quality—in a single scalar, δ_{jmt}. This transformation has five unknown compo-

[26] The solution method's complexity is proportional to the product of the dimensions of quality, rival firms, quality shocks, and heterogeneity in consumer quality preferences. Thus, adding moderate heterogeneity can increase the time required to solve this problem from months to years. However, the method can solve the scoring design problem with heterogeneous quality preferences with fewer firms or quality dimensions.

[27] 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.

nents: preference heterogeneity $(\tilde{\alpha}_i, \tilde{\beta}_i)$, preferences for TM $(\boldsymbol{\lambda}^d)$, switching costs $(\boldsymbol{\lambda}^l)$, and plan-market-year fixed effects $(\boldsymbol{\delta})$. Collecting these in a vector $\boldsymbol{\vartheta}$, the first stage solves

$$\max_{\boldsymbol{\vartheta}} \quad \underbrace{\sum_t \sum_i w_{it} \sum_{j \in \mathcal{J}_{m(i)t}} y_{ijmt} \ln(s_{ijmt}(\boldsymbol{\vartheta}))}_{\text{weighted log-likelihood}} \quad \text{s.t} \quad \underbrace{s^*_{jmt} = \sum_i w_{it} s_{ijmt}(\boldsymbol{\vartheta})}_{\text{share matching}} \quad \forall j, m, t \quad , \tag{9}$$

where y_{ijmt} is a choice indicator, $s_{ijmt}(\boldsymbol{\vartheta})$ is the model-implied individual choice probability, and s^*_{jmt} is the observed market share. Thus, the first step is a constrained weighted maximum likelihood problem, where w_{it} are nationally representative MCBS sampling weights. The constraint matches predicted and observed market shares, which I solve using the Berry (1994) inversion and the Berry *et al.* (1995) fixed-point contraction.

The second step is a two-stage least-squares regression of the estimated mean preferences on their components. I decompose consumers' unobserved preference (ξ_{jmt}) into systematic taste for MA plans in each market $(\mathbb{d}_{mt})$, preferences for the contract-year $(\eta_{c(j)t})$, and all residual unobserved preference $(\tilde{\xi}_{jmt})$. Since scores are assigned at the contract-year level, $\eta_{c(j)t}$ also absorbs consumers' expected utility from quality.

$$\hat{\delta}_{jmt} = \underbrace{\alpha p_{jmt}}_{\text{premium}} + \underbrace{\beta b_{jmt}}_{\text{benefits}} + \underbrace{\boldsymbol{\lambda}^{a\prime} \boldsymbol{a}_{jmt}}_{\substack{\text{plan}\\ \text{attributes}}} + \underbrace{\eta_{c(j)t}}_{\substack{\text{contract-year}\\ \text{FE}}} + \underbrace{\mathbb{d}_{mt}}_{\substack{\text{market-year}\\ \text{FE}}} + \tilde{\xi}_{jmt} \quad \forall j, m, t \tag{10}$$

Firms' knowledge of $\tilde{\xi}_{jmt}$ when pricing renders premiums and benefits endogenous in this regression. To address this issue, I develop two instruments based on regulatory features of insurers' additional revenue $(R(\cdot))$. First, I use an average of TM's insurance cost in the plan's other markets. The regulation links each plan's subsidies with the public option's cost in every county in which it participates, making the leave-one-out average a strong predictor of subsidies unaffected by local demand. Second, I use variation across plans in the added revenue they obtain when pricing below the kink. This second instrument helps distinguish between the effect of endogenous prices on premiums and benefits. Both instruments vary across plans and years due to variations in county choices, regulations, and TM's cost.[28]

Consumers' unobserved preferences for products also influence contract quality. Consumers, however, do not observe plan quality but its discretization in scores. As only variation within a contract over time will be used next to decompose $\eta_{c(j)t}$ into consumers'

[28] The exclusion restriction would fail if, for example, plans change counties due to the correlation between TM cost and plan-specific preference. As 92% of non-terminated plans remain in a county the following year, this concern seems unlikely. The second instrument corresponds to the rebate fraction for plans pricing above the benchmark and one for the rest. See Online Appendix Table 3 for first-stage estimates.

preferences for score-years and contracts, the only source of endogeneity is from variation over time in unobserved plan preferences. For this residual to affect the estimates, it must dominate the effect of contemporary changes in cost, design, and investment risk. Online Appendix II shows that using instruments that rely only on variation in scoring design and firms' investment costs to address this residual endogeneity does not meaningfully change the estimates. Therefore, I do not instrument for score-years in the main estimates, using instead the variation in design to separately identify consumers' preferences and beliefs about quality from their valuation for scores, as will be described next.

VI.A.1 Quality beliefs and preferences: In estimation, consumers' preferences for scores are star-year fixed effects absorbed within $\eta_{c(j)t}$. Their separate identification from variation in the score assigned to a plan follows standard demand identification results (Berry and Haile, 2020). Intuitively, consumers reveal these preferences when trading off premium increases for star-rating changes. The challenge is that these valuations do not reveal consumers' preferences for quality separately from their beliefs. For example, consumers might be willing to pay a substantial amount for plans to have 4 instead of 3 stars, all else equal. This observation, however, can be based on a belief that four-star plans are of starkly superior quality or because consumers substantially value even small differences in quality. Disentangling beliefs from preferences requires an assumption on how consumers form beliefs, given the scores they observe and how these vary in the data. I consider two juxtaposed assumptions regarding beliefs:

Assumption 1 (Consumer beliefs)**.** *One of the two hold: (1)* ***Informed choice****: Consumers know $\psi_t(\cdot)$ and use scores and Bayes' rule to update a continuous prior density $f : \mathcal{Q} \to \mathbb{R}_+$, with compact and connected support; (2)* ***Ignorance****: Consumer's posterior beliefs $\mathcal{E}[\boldsymbol{q}|r, \psi]$ are exogenous, independent of ψ, and bounded in $\mathcal{Q}$.*

Informed choice is the most common assumption in the literature. In theoretical work, consumers (receivers) often know precisely the rules by which the regulator (sender) transforms the distribution of quality (state) (Kamenica and Gentzkow, 2011). In the empirical literature, consumers either know the true structure or a parametric and unbiased approximation of it (Crawford and Shum, 2005; Dranove and Sfekas, 2008; Barahona *et al.*, 2022).[29] Crucially, in both cases, the econometrician knows how consumers interpret scores and can rely on their variation. In addition, consumers' knowledge of the scoring rule allows the regulator to shape their beliefs, which gives substantial power to the scoring policy; the

[29] Alternatively, some allow for parametric bias based on additional data, such as external surveys.

ignorance assumption generates the other extreme of weak regulatory power.[30]

These assumptions gain power once combined with appropriate variation in scores. In Online Appendix II, I show that since MA scores are partitions of weighted averages of quality, they are well approximated within the class of *monotone partitional scores* (Dworczak and Martini, 2019). This class includes all scores that partition the space of quality (e.g., the five-dimensional space of average category-level quality) into numbered partitions, assigning weakly greater labels to strictly greater quality.[31] Therefore, to score a plan, one needs only to assess in which partition its quality falls. This class of scores is exceedingly common and includes all deterministic certifications of quality (e.g., front-of-package nutrition labels), letter grades (e.g., restaurant hygiene scores), and many others.

Assumption 2. *(Design variation) ψ_t is drawn from a distribution with a strictly positive density over partitional scores with linear boundaries and $N \geq 3$ partitions, with N fixed.*

Assumption 2 states that scores will continue to vary within a set that includes, but is not limited to, the type of designs observed in the data. It does not require that the number of partitions grow with the sample or entail complex aggregation rules (i.e., boundaries). The key identification result, proven in Online Appendix II, follows.

Proposition 1. *(Quality beliefs and preference identification) Let assumption 2 hold and quality preferences be linear, i.e., $v(q) = \boldsymbol{\gamma}'\boldsymbol{q}$. If assumption 1.a holds then $(\boldsymbol{\gamma}, f(\cdot))$ are identified. If assumption 1.b holds, then there is nontrivial identified lower bound for $\boldsymbol{\gamma}$.*

This general identification result depends on the setting only insofar as common consumer preferences for score-years can be identified, and the scoring design varies within a typical class.[32] Intuitively, 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 $[0, 1]$, consumers would be willing to pay $8/9$ more for a top-rated product than a bottom-rated one. Some simple

[30] Alternatively, consumers could hold rational expectations about quality. This would imply that they know scoring rules, firms' cost structure, and investment risk well enough to predict endogenous quality change. Hence, informed choice is a relaxation. Moreover, rational expectations allow the regulator to control quality without informational losses, which renders informational policies even stronger than considered here.

[31] Dworczak and Martini (2019) consider a larger class of monotone partitional signals that allow for the full revelation of quality in some partitions. The order used also varies across applications.

[32] The result does not depend on the logit structure. Moreover, the restriction to linear partitions is immaterial. The full support assumption on scoring design is valuable to identify the corners of prior beliefs. Still, the identification argument is not at the limit: Variation in design provides meaningful identifying restrictions even if all partitions have positive measures.

algebra shows that by observing the differences in willingness to pay and knowing the scoring structure, we can bound γ within $(8/9, 8/7)$. Scoring variation produces new intervals for γ, which intersect and shrink the identified set down to a point. This process also bounds posterior beliefs and thus priors.

The result shows that informed choice is a powerful assumption. It imposes a strong structure on consumers' understanding and, in return, delivers identification. In MA, this assumption mostly requires that consumers know the relative contribution of categories to the scores. In Online Appendix I, I provide evidence to support this assumption and reject the ignorance conjecture. Therefore, I rely on the informed choice assumption for the primary analysis and return to the case of ignorance in the final section for robustness.

For results that rely on informed choice, I estimate preferences (now captured by a vector $\boldsymbol{\gamma}$) and prior beliefs ($f(\cdot)$) using a nonparametric minimum distance estimator. To remove any systematic preference for specific contracts, I only leverage time-series variation in consumers' preferences for contract-years, $\eta_{c(j)t}$. The resulting estimator is

$$\min_{\boldsymbol{\gamma},\boldsymbol{\zeta}} \sum_{c(j)} \sum_{t} \sum_{\tau>t} \left(\Delta_t^\tau (\eta_{c(j)t} - \boldsymbol{\gamma}' \mathcal{E}[\boldsymbol{q}|r_{c(j)t}, \psi_t; \boldsymbol{\zeta}])\right)^2 \quad , \tag{11}$$

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

VI.A.2 Estimates: Table 1 presents the main demand estimates.[33] Panel A shows the estimated preferences for premium and benefit levels. A dollar in benefits is roughly equivalent to a \$2.5 reduction in premiums for a low-income male of "fair" perceived health.[34] 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.[35] As I will show later, this elasticity implies reasonable markups for

[33] Online Appendix Table 4 presents the remaining estimated coefficients.

[34] The discrepancy between my finding and those of Abaluck and Gruber (2011) are, in part, due to \$1 in benefits translating to less than a \$1 reduction in expected spending, which inflates the coefficient. I discuss this further in Online Appendix II.

[35] 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. The table also displays premium elasticities comparable to those of Miller *et al.* (2022). Their estimate is -2.6 using similar data but a different model. In my model, this premium elasticity would imply excessive price elasticities and negligible firm markups.

Table 1: Key Demand Estimates

Panel A:	**Premium** (α_i)		**Benefits** (β_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 ≥ 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 (λ^a)			**Panel C: Quality preferences** (γ)		
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)

Observations	36,447	Weighted log. likelihood	-5.131
Mean price elasticity	-8.348	Mean premium elasticity ($p^C > 0$)	-0.951

Notes: Panels A and B report estimates of preference as in equation (4). In Panel A, the omitted category is low-income males of "fair" self-reported health status. Income groups are terciles of the MCBS 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. Unadjusted heteroskedastic standard errors in parentheses. *p<0.05, **p<0.01, ***p<0.001.

firms in this market.

Panel B presents some of the additional preferences for fixed product attributes. Importantly, consumers strongly prefer products that bundle prescription drug coverage. Panel C of Table 1 presents consumers' quality preferences under the assumptions of informed choice and linear quality preferences. The most valued quality category is Medical Outcomes (i.e., provider quality), with consumers' willingness to pay for maximal quality in the category being $4,498 in yearly premiums. The least valued category is Intermediate Outcomes (i.e., chronic condition management), with a maximal willingness to pay of $1,654. However, quality dispersion differs across categories, and a standard deviation increase is worth $1463.2 in Medical Outcomes and $204.4 in Intermediate Outcomes.

Following Train (2015), I compute the surplus loss from consumers' incomplete information about quality, holding product attributes fixed. The average consumer loses \$185.9, or three monthly premiums, per year due to two informational frictions. First, *within scores*, the quality of products is indistinguishable. For example, the average spread in quality between the best and worst 4-star plans is equivalent to a \$257.1 difference in premiums. Second, *across scores*, misalignment between consumers' preferences for quality categories and their relative contribution to the score makes it such that higher-scoring products can have lower quality-utility than lower-scoring ones. On average, 22.4% of plans have a lower-scoring alternative that delivers a higher quality-utility. Thus, although consumers and CMS agree that plans of higher overall quality should receive higher scores, they disagree on how to weigh different dimensions. Decomposing the total surplus loss into these two factors reveals that 95% stems from across-score frictions.[36] Within-score frictions are limited by firms' incentives to target the lower boundaries of scores and by the small number of equally scored alternatives offered in each county.

VI.B Supply

VI.B.1 Insurance marginal costs: Insurers' pricing first-order optimality condition equates marginal revenue with marginal costs.[37] Since revenue depends only on observed demands, prices, and estimated elasticities, this condition can be used to recover the marginal cost parameters ($\boldsymbol{\theta}^c$). Assuming marginal costs are linear, the resulting condition is

$$\underbrace{\boldsymbol{p}_f + R(\boldsymbol{p}_f, \boldsymbol{z}_f)}_{\text{revenue per consumer}} + \underbrace{(\nabla \tilde{\boldsymbol{D}}_f')^{-1}(I + \nabla R_f(\boldsymbol{p}_f, \boldsymbol{z}_f))\tilde{\boldsymbol{D}}_f}_{-\text{profit margin}} = \underbrace{\boldsymbol{\theta}_q^{c\prime}\boldsymbol{q}_f + \boldsymbol{\theta}_a^{c\prime}\boldsymbol{a}_f + \boldsymbol{c}_f}_{\text{marginal cost } = \boldsymbol{C}(\boldsymbol{q}_f, \boldsymbol{a}_f, \boldsymbol{\theta}^c)}, \tag{12}$$

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. On the right-hand side, I have decomposed the firm's marginal cost into its quality components ($\boldsymbol{q}_f$), systematic observable components ($\boldsymbol{a}_f$), and residual ($\boldsymbol{c}_f$).

Variations in demand, competition, and regulation identify marginal costs. The first column of Table 2 presents the estimates of $\boldsymbol{\theta}_q^c$ when $\boldsymbol{a}_f$ includes contract, year, and market fixed effects and controls for bundled services. Quality's effect on marginal costs is identified by the residual correlation between marginal revenue and quality after accounting for market and national quality trends. The estimates indicate that improving both types of medical

[36] This is done by simulating a scenario without within-score frictions: Consumers first choose a plan based on expectations and then get to adjust their choice among plans of the same score with full information.

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

Table 2: Quality's Insurance Costs, Investment Costs, and Marginal Welfare

	Investment Cost ($\boldsymbol{\theta}_q^c$)		Insurance Cost ($\boldsymbol{\mu}_k$)		Marginal Welfare	
Access	31.160	(16.690)	15.620**	(5.965)	21.421	[-41.5, 101.3]
Intermediate	108.400***	(12.800)	19.530***	(4.963)	-108.123	[-165.3, 0.7]
Outcome	16.810***	(3.832)	15.000*	(6.516)	119.795	[57.2, 165.0]
Patient	-244.300***	(57.540)	14.730*	(7.424)	86.078	[-5.5, 151.1]
Process	-175.600***	(27.560)	1.106	(4.718)	-32.423	[-94.3 51.1]
	N $28,966$	R^2 0.531	N $5,281$	R^2 0.261	N $1,697$	

Notes: This table reports the estimates of $\boldsymbol{\theta}_q^c$ in the marginal cost equation (12), $\boldsymbol{\mu}_k$ in the investment cost equation (14), and the derivative of total welfare at expected quality. Values in the left column are in dollars per member-month and in the rest they are in millions per contract-year. The sample size in the last column is smaller due to the intersection with the MCBS counties. Standard errors in parentheses are heteroskedasticity robust. 25th and 75th percentiles in square brackets. *p<0.05, **p<0.01, ***p<0.001.

outcomes increases marginal cost. In contrast, improvements in Process and Patient quality lower marginal costs. For Process, this reversal is likely due to its effect on preventive care and managing expensive chronic illnesses, which can prevent expensive hospitalization (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. Nevertheless, these improvements might come at the expense of large investment costs.

These estimates imply reasonable markups for insurers, with an average of 11.2% and 13.3% for the top four insurers. Using claims data for the top insurers during 2010, Curto *et al.* (2019) estimate an average cost of \$590 per enrollee risk-month in medical costs, or \$680 in adjusted 2015 dollars. My estimate for the same set of firms is \$771, including 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.

VI.B.2 Investment costs: Since investments are unobserved, marginal investment costs cannot be inferred pointwise from optimality conditions, unlike insurance costs. However, the distribution of investment risk can be estimated using standard deconvolution results given observed quality realizations (Schennach, 2016). Online Appendix II details the formal identification and nonparametric estimation procedure for risk. Using the distributions of risk and quality, I compute the likelihood of investments, which allows me to evaluate the optimality conditions in expectations. Therefore, the combination of optimality and the distribution of risk and quality can be used to identify the remaining cost parameters.

Formally, the first-order condition of investment for firm f in category k in year t equates marginal revenue ($\frac{\partial}{\partial x_{ckt}}\mathbb{E}[V_f(\boldsymbol{q}_f, \boldsymbol{q}_{-f}, \psi_t)|\boldsymbol{x}_{ft}]$) with marginal investment cost ($\frac{\partial I(\boldsymbol{x}_{ft},\mu_{ft})}{\partial x_{ckt}}$). Given observed quality $\boldsymbol{q}_{ft}$, I decompose the marginal revenue into its conditional mean and variance. In the optimality condition, this decomposition gives the regression equation[38]

$$\mathbb{E}[\frac{\partial}{\partial x_{ckt}}\mathbb{E}[V_f(\boldsymbol{q}_f, \boldsymbol{q}_{-f}, \psi_t)|\boldsymbol{x}_{ft}]|\boldsymbol{q}_{ft}] = \frac{\partial I(\boldsymbol{x}_{ft}, \mu_{ft})}{\partial x_{ckt}} + \nu_{ckt} \qquad \mathbb{E}[\nu_{ckt}|\boldsymbol{q}_{ft}] = 0 \quad . \tag{13}$$

To operationalize this regression, I assume firms' investment costs are quadratic and separable across products and categories, $I(\boldsymbol{x}_{ft}, \boldsymbol{\mu}_{ft}) = \sum_{c\in\mathcal{C}_f, k\in\mathcal{K}} \mu_k(x_{ckt} - \underline{x}_{kt})^2 + \mu^F_{fkt}(x_{ckt} - \underline{x}_{kt}))$, 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 results in

$$\underbrace{\mathbb{E}[\frac{\partial}{\partial x_{ckt}}\mathbb{E}[V_f(\boldsymbol{q}_f, \boldsymbol{q}_{-f}, \psi_t)|\boldsymbol{x}_{ft}]|\boldsymbol{q}_{ft}]}_{\substack{\text{conditional expectation}\\ \text{of marginal insurance profits}}} = \underbrace{2\mu_k(\Phi_k^{-1}(q_{ckt}) - \Phi_k^{-1}(\underline{q}_{kt})) + \mu^F_{f(c)kt}}_{\substack{\text{conditional expectation}\\ \text{of marginal investment cost}}} + \tilde{\nu}_{ckt} \quad . \tag{14}$$

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 investments' space. Assuming that this second error is mean-zero conditional on $\boldsymbol{q}_{ckt}$, equation (14) is a linear regression.

I estimate marginal investment costs using OLS. To avoid mixing contracts with distinct cost structures, I limit attention to HMO and PPO contracts, which account for 81% of enrollment, and estimate costs separately across states.[39] The second column of Table 2 displays the estimated coefficients and shows that Intermediate Outcome quality is the most expensive to improve, while Process measures are the cheapest. In comparative terms, the estimates suggest contracts in the 75th percentile of Access quality invested an average of 5.16 million more than those in the 25th percentile. Overall, the median-quality firm invests 24.6% of its insurance profits.

[38] In Online Appendix II, I show that the left-hand side of this expression is a function of only identified distributions and has an analytical expression.

[39] This 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. 18% of beneficiaries choose multi-state contracts, and the median multi-state contract has 80% of its population in a single state.

Using the estimates, I evaluate the efficiency of quality provision by computing the marginal welfare value of quality.[40] Column three of Table 2 shows that for the average contract, a marginal increase in its Access, Outcome, or Patient qualities would increase consumers' surplus by more than it would cost to produce, and the opposite holds for Intermediate and Process. Hence, on average, the first three categories are underprovided and the other two overprovided. Aggregating quality dimensions according to consumers' preferences indicates that aggregate quality is underprovided. However, 33% of contracts overprovide quality, which is more common among 5-star plans (41.2% overprovide) and those below 2.5 stars (49.2% overprovide). Five-star plans invest excessively in categories overweighted by the scores relative to consumers' preferences. Low-quality plans have high investment costs overall and inefficiently steer demand away from alternatives. The following section shows that these findings are partially due to the spencian distortion and the categories' contributions to scores. The former distorts aggregate provision, while the latter distorts the relative allocation of investments across quality categories.

VII Scoring Design

In this section, I use the model to specify and solve the regulator's optimal scoring design problem under the assumption of *informed choice.* I find the optimal design within the monotone partitional class and decompose its regulatory mechanisms. I discuss optimal scoring granularity, multidimensional quality aggregation, competition, and regulatory distortions.

VII.A Objective and approach

The designer seeks to maximize the expected sum of consumers' surplus and insurer's profit, subtracting the government's spending on enrollment subsidies:[41]

$$\max_{\psi\in\Psi}\int[\underbrace{CS(\psi,\boldsymbol{q})}_{\substack{\text{Consumer}\\\text{surplus}}}+\rho^F\underbrace{\sum_f V_f(\psi,\boldsymbol{q})-I(\boldsymbol{x}_f^*(\psi),\mu_f)}_{\substack{\text{Insurer}\\\text{profit}}}-\underbrace{\rho^G Gov(\psi,\boldsymbol{q})}_{\substack{\text{Government}\\\text{spending}}}]dF(\boldsymbol{q}|\boldsymbol{x}^*(\psi)) \tag{15}$$

Above, the designer chooses a scoring rule ψ from a class Ψ, acknowledging its effect on equilibrium investments, prices, beliefs, and enrollment choices. This endogenous invest-

[40] I compute this as the derivative of the sum of expected consumer surplus and insurer profit with respect to quality. I evaluate it at the observed quality and hold prices fixed.

[41] I evaluate consumers' realized surplus (Train, 2015). That is, $CS(\cdot) = \sum_i \ln(\sum_{j\in\mathcal{J}_{m(i)}} \exp(\delta_{ij})) + \sum_{j\in J_{m(i)}} s_{ij}(\delta^*_{ij}-\delta_{ij})$, where δ^*_{ij} is the consumer's realized utility given true instead of expected quality.

ment response presents the designer with a trade-off: For any fixed investment distribution, more information helps consumers choose and makes competition more effective. However, firms might invest inefficiently under full information—a distortion coarser information can regulate. Therefore, the scoring rule must trade off information for efficiency.

Solving this trade-off is challenging. First, scoring rules are discontinuous mappings from quality space down to a few scalars. There are no known optimality conditions, and a priori, the loss from approximations is unbounded. Second, because the regulator computes an expectation over quality, evaluating designs requires integrating over a continuum of counterfactual subgame equilibria. I draw on two insights to address these challenges and develop a method to divide the design problem into a series of smaller, manageable ones.

First, I show in Online Appendix III that any monotone partitional score is a composition of a polynomial *aggregator*, which aggregates multidimensional quality into an index, and a *cutoff* function, which partitions the index into scores. Therefore, we can find the optimal design by finding the best one for all subproblems constrained to a particular number of cutoffs and aggregator polynomial order (i.e., the boundary curvature). Each of these problems is moderately simple, conditional on being able to compute the regulator's integral.

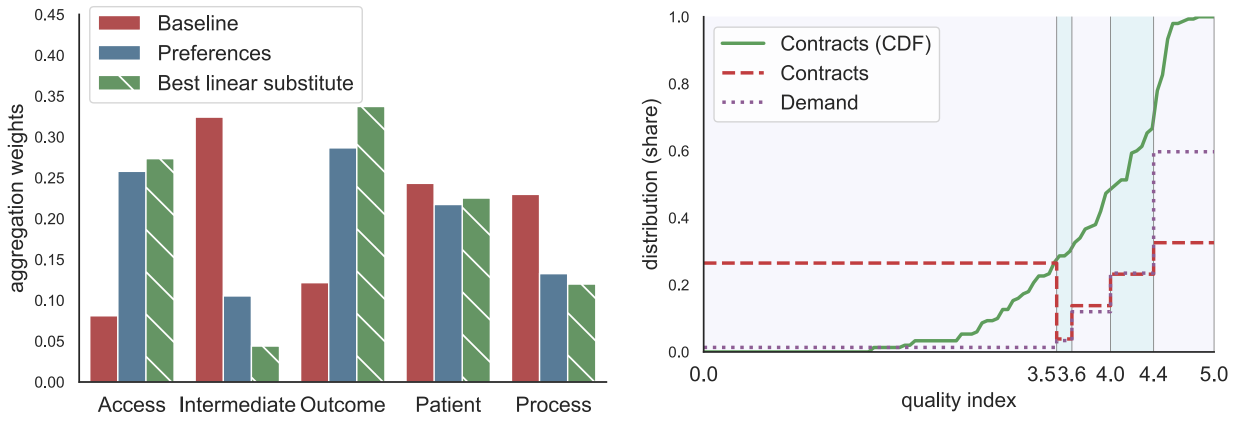
The second insight addresses the integral and comes from Aumann *et al.* (1995), who note that selecting a disclosure policy is akin to choosing a distribution of posterior beliefs. In scoring design, the analogous statement is that each score generates a distribution over qualities, score valuations ($\mathcal{E}[\boldsymbol{\gamma}'\boldsymbol{q}|r,\psi]$), 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 these evaluations. Therefore, the integral of any policy is a known weighted sum of points in the grid.

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 $\rho^G = 0$.[42] I show results for other objectives in Online Appendix III. The central mechanisms and qualitative results remain the same.

VII.B Optimal constrained design

Figure 5 shows the optimal design constrained to scores with at most fifteen partitions and quadratic aggregators. The resulting solution, however, has a linear aggregator and uses only five scores; I call this solution the *best linear substitute*, as it can be implemented through minimal adjustments to CMS's current rules. Three key features generate this policy's market effects: the lowest scoring segment in the cutoff policy is broad, creating a *pooling at*

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



(a) Aggregation weights

(b) Quality distribution and cutoffs

Figure 5: Best Linear Substitute Design

Notes: Figure (a) compares the optimal aggregation weights with CMS's scheme and consumers' preferences. Figure (b) shows the cutoff placements, with segments of changing colors indicating different scores.

the bottom effect; the cutoff function has limited granularity, using only five scores; and the aggregator is better aligned with consumers' preferences, placing relatively higher weights on quality dimensions they prefer. Next, I discuss these features and their key mechanisms.

VII.B.1 Pooling at the bottom: The first stark difference between the new design and the Star Ratings is how they classify low-quality plans. In the baseline, the Star Ratings partition the quality space uniformly, which allows consumers to distinguish between low and moderately low-quality plans. While this information is valuable to consumers, it distorts quality provision. By pooling at the bottom, the new design uses the within-score informational friction to induce low posterior beliefs among consumers for low-scoring plans; this shifts their demand toward better scores. This shift in demand incentivizes higher aggregate investment and remedies the aggregate underprovision problem documented above.

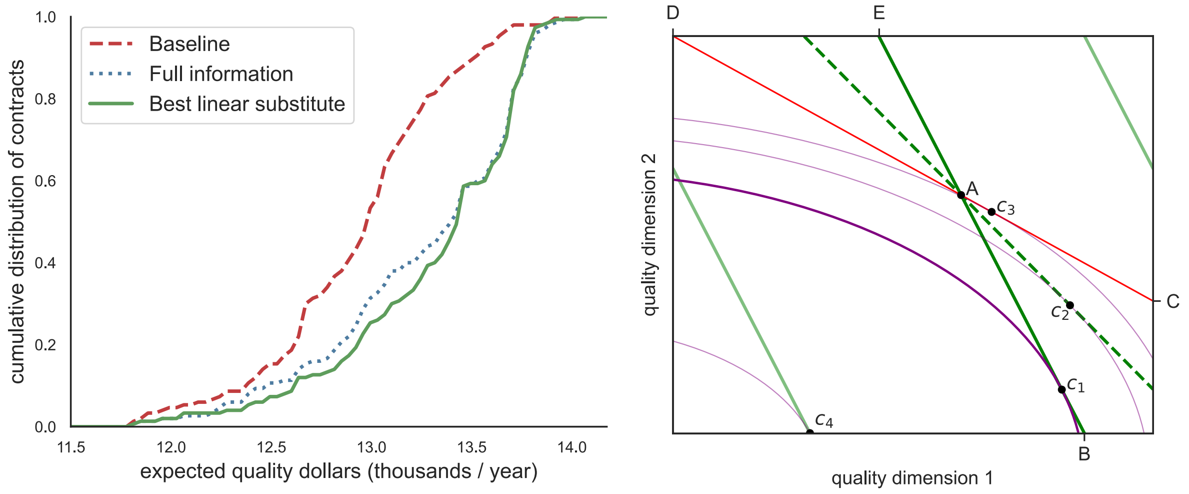
As the range of higher qualities is more finely partitioned than in the baseline, consumers' counterfactual choice is among higher-quality products with greater information about them. The mechanism is the same as in Section II, which pools quality to deal with underprovision. In the status quo, 62.6% of contracts would receive 1 star under the new design and enroll 18.9% of MA consumers. In equilibrium, this low demand stimulates investments such that only 26.5% of contracts obtain the lowest score and only 1.3% of consumers enroll in them. These contracts are virtually exiting the market since their investments are minimal (bounded by minimum standards). In contrast, top-scoring plans account for 32.6% of supply and serve 59.7% of consumers. Figure 6a shows the distribution of quality under a full-information counterfactual and demonstrates that the left tail of quality in the regulated market shifts inward as incentives to underprovide quality are alleviated.

VII.B.2 Limited granularity: The granularity of scores equals the number of potential investments firms might consider optimal. Intuitively, without investment risk, firms will always aim to be at one of the cutoffs since any interior investment does not translate to increased demand. This observation is known as the *delegation equivalence* of scores (Kolotilin and Zapechelnyuk, 2019), which helps explain the effect of scoring granularity on the supply of heterogeneous products. The counterpart to this supply is consumers' heterogeneity in willingness to pay for quality. As preference heterogeneity grows, so does the optimal heterogeneity of products. Hence, a more granular scoring system allows products of different qualities and prices to match with consumers of different tastes. The trade-off, however, is that firms' incentives to provide quality suffer from Spencian distortion, and as their production flexibility grows, these distortions increase. Hence, there is only one optimal cutoff in a setting of homogeneous preferences and firms since there is a unique optimal quality level. In contrast, the optimal granularity is infinite in an environment with heterogeneous consumers and firms but no Spencian distortion. In MA, the optimal number of scores is 5—four fewer than the current system, or as in an A-F letter-grade system. Any more and the loss from quality distortion exceeds the surplus gains from variety.

VII.B.3 Aggregation weights: The final and, as will be discussed next, most important feature of the new design is its aggregation weights. In quality space, these weights determine the slope of the boundaries that separate one score from the next. According to the model, consumers' indifference curves for quality lie in the same space as straight lines. Figure 6b illustrates this in the case of two-dimensional quality, with line BE being the boundary between the second and third score and line DC consumers' indifference. The new design improves the alignment between boundaries and indifferences by placing relatively more weight on the categories consumers value most, transforming BE into the dashed line. This change ameliorates two failures caused by quality aggregation.

The first loss stems from a regulatory moral hazard problem. Ignoring investment risk, firms first choose which score their plans should have and then find the cost-minimizing way to attain such a score.[43] For example, in Figure 6b, point c_1 marks the tangency of a firm's isocost curve (purple) with the scoring threshold (green), which would be the efficient investment combination for it to attain the third score. For the regulator, however, this decision ignores consumers' preferences over the relative allocation of quality, which translates into a

[43] The same ideas apply with investment risk but in expectations. Firms tend to invest above the thresholds since quality enters through demand, the concavity of which induces risk attitudes in their behavior.



(a) Quality underprovision

(b) Weight alignments

Figure 6: Scoring design mechanisms

Notes: Figure (a) plots the distribution of quality in the baseline, full-information benchmark, and under the best linear substitute design. Figure (b) illustrates the aggregation mechanism with two quality dimensions and 4 scores. The line EB is the second scoring cutoff, and DC is the consumers' indifference curve. Products above EB obtain a score of 3, and below it, 2. The misclassification region is $DEA + ABC$, since consumers prefer products in the former over the latter. The dashed green line represents the new design. Purple concave lines represent a firm's isocost curve under different total investment levels.

multitasking moral hazard problem (Holmstrom and Milgrom, 1991). This friction explains the mixed provision problem documented in Section VI. Improving the alignment of boundaries and preferences alleviates this problem, which renders firms' incentives to substitute investments across quality dimensions similar to consumers' marginal rate of substitution.

The second loss from aggregation is the across-scores informational distortion, visible in Figure 6b since products in the triangle ΔDEA are preferred by consumers to the higher-scoring ones in ΔABC. The optimal partial alignment of boundaries and preferences narrows the misclassification region, reducing across-score frictions by 97.1% in expectation. In total, the mean squared error of consumers' beliefs regarding plan quality drops by 91.7%.

The optimal design does not feature perfect alignment due to heterogeneity in investment costs. Since heterogeneous firms sort themselves along the scoring boundaries, changes in alignment force firms to adjust their overall investment. Discontinuities across scores imply that these adjustments can lead to substantial losses in overall quality production and welfare. For example, in Figure 6b, as EB rotates around point A, the firm that formerly invested in c_1 must now invest in a higher level c_2 to obtain the same score. As investment costs are convex, further tilting might dissuade the firm from maintaining a score of 3 at

Table 3: Counterfactual welfare changes

	ΔSurplus	ΔProfits	ΔWelfare	ΔGov.	%MA
Panel A: Expected welfare (informed choice)					
Best linear substitute	146.5	522.8	669.2	-76.4	57.1%
↪ Information only	44.8	213.7	258.8	-63.5	45.0%
↪ Information and prices only	-36.7	270.8	234.1	-10.2	41.0%
Full information	136.0	488.6	624.6	-69.1	54.4%
↪ Information only	129.5	298.3	427.9	-109.4	52.9%
↪ Information and prices only	-5.3	374.8	369.5	-48.1	47.7%
CMS-weighted	93.1	164.6	257.7	-81.6	41.7%
Best linear certification	151.8	480.5	632.4	-88.4	56.0%
Panel B: Robust design minimum welfare (ignorance)					
Robust linear substitute	5.6	288.6	294.3	-154.6	50.7%
Best linear substitute	-344.4	400.9	56.5	-148.3	56.6%
CMS-weighted	72.1	173.0	245.1	-107.9	42.0%

Notes: This table displays welfare changes in 2015 dollars per Medicare beneficiary relative to the status quo. Panel A presents expected welfare under the assumption of informed choice and identified consumer preferences. Subitems show results holding quality at the baseline level, isolating the counterfactual effects. Panel B presents the worst-case welfare under the assumption of ignorance and set-identified preferences, as studied in Section VIII. Government spending accounts for subsidy and rebate payments, including TM. The baseline simulated market share of MA is 27.8%.

point c_3, pushing it to a score of 2 at c_4. Evidently, the firm's new quality at c_4 is substantially lower than at c_2. Thus, the optimal weights allow for some inefficiencies in relative investments and information to maintain aggregate quality at desired levels. Overall, for plans not in the bottom score, quality increases in all categories except in the overprovided Intermediate Outcomes, which falls by 1.2%. Process quality increases despite its overprovision and lower weight since it is uniformly cheap to produce. This allows firms with high Outcome investment costs to attain moderate scores and retain demand.

VII.C Welfare

The first row of Table 3 shows the estimated welfare gains from replacing MA Stars with the best linear substitute design. Per Medicare beneficiary, the alternative increases surplus by $146.5, or 2.4 monthly premiums (Parts C + D), and profits by $522.7. Average contract premium and markups drop by 45.8% and 1.5%, respectively, while aggregate quality increases by 2.5% (in premium-equivalent dollars, according to consumers' preferences). However, scores shift consumers' demand toward higher-quality plans, which increases the premium paid per consumer by 125.2% and the markup on chosen plans by 37.5%. Chosen quality is 4.3% higher. In total, firms' profit per enrollee grows by 79.5% which, compounded by a large substitution from TM to MA, increases total firm profits by 271.8%.

Consumers switch from TM to MA as quality and information improve, allowing them to benefit from MA's generous cost-sharing. Consumers' preference for TM can be seen as partially driven by the insurance value of broad networks against unexpectedly bad providers. Better information and quality under the counterfactual system offset the quality risk induced by MA's narrower networks.[44] Consumers who switch to MA often choose plans that cost less to subsidize than TM, which decreases spending by 0.8% per Medicare beneficiary.

VII.C.1 Asymmetric information and moral hazard: The alternative design changes information, quality, and prices. It alleviates frictions due to asymmetric information and firms' moral hazard and changes the degree of differentiation across firms, which affects market power over prices. To assess these changes, I compute equilibria holding prices and qualities fixed at their baseline level under both the best linear substitute design and a fully informative scoring scenario.

Table 3 shows that information alone accounts for 38.7% of the welfare impact of the alternative scoring system. The new system is more informative, increases surplus, and expands enrollment in MA. However, coarse scoring is purely distortive when prices and qualities are exogenous: Welfare under full information is 65% higher.[45] Allowing prices to adapt to the new information structure while holding quality constant allows firms to capture surplus from vertical differentiation. Welfare under the best linear substitute design drops by $24.7 per beneficiary because efficiency is lost to market power. Firm profits increase, but less than the surplus loss as consumers substitute back to TM. Coarsening consumers' information is still distortive in this case, and full information welfare is 58% higher. However, ignoring quality changes, consumers' surplus is slightly higher under CMS's design than in full information. The baseline scoring system limits firms' vertical differentiation and, therefore, their ability to extract surplus from consumers.

Allowing for equilibrium quality responses overturns the dominance of full information.

[44] Similar counterfactual market expansion effects have been predicted for optimal subsidy design in this market (Miller *et al.*, 2022). For comparison, MA grew from 27% to 48% between 2012 and 2022.

[45] The information-only effect under full information is analogous to the surplus loss computed in the demand estimates. The slight difference is due to the previous averaging distortions from 2009 to 2015.

Total welfare under the alternative design increases by 285% as quality adapts to the optimal scoring boundaries and incentives, exceeding full-information welfare by 7%. As noted above, full-information outcomes are subject to inefficient underprovision due to Spencian distortion. Welfare values corroborate that the shift in quality produced by the best linear substitute (Figure 6a) improves welfare. Low willingness-to-pay consumers benefit from higher quality in the lower segments, and firms benefit from a coordination effect that leads to market expansion.

Accounting for moral hazard in quality provision changes the optimal design, the desirability of informativeness, and the principal channel through which scores affect the market. The results and mechanisms reveal a simple intuition: Scores are inherently informative because quality bunches at boundaries known to consumers. Therefore, the main design challenge is positioning cutoffs to dictate quality rather than alleviating informational asymmetries. Boundaries, however, must be readily interpretable by consumers, which means that they can be mapped to a narrow set of utilities. Across-score frictions in the Star Ratings limit interpretability, which is why information can still be improved.

The gap between the constrained optimal design and the full-information outcome has two meaningful implications for policy design. First, the ability to approximate full-information outcomes with simple coarse scores is valuable in settings where the underlying quality data are complex or subject to privacy regulations. For example, regulators might be unwilling to disclose the detailed performance of small insurers since others might use it to identify their populations and discriminate against them. Yet, as in the example of Section II, consumers in a scored market can behave as if fully informed, even if they cannot detect large deviations in quality by firms. Second, the gap implies that even if consumers are highly sophisticated Bayesian agents, they might still benefit from coarse information. Thus, the optimal design need not conflict with behavioral concerns about the ability of enrollees to process complex information; it is not the case that sophisticated consumers prefer complex signals of quality.

VII.C.2 Decomposition by design feature: The new design improves on the existing system by adjusting weights and cutoffs. To understand the role of optimal granularity and cutoff location, I solve for the optimal design holding CMS's weighting system fixed. Table 3 shows the resulting welfare.[46] The net change in optimizing cutoffs is positive for both consumers and firms, which indicates that even if CMS has external reasons for prioritizing specific categories, the current cutoff design can be improved. This reoptimized system attains 35% of the welfare gains of the optimal design.

[46] Cutoff locations are illustrated in Online Appendix III.

To isolate the effects of optimal weighting, I solve for the optimal dichotomous certification of quality. Table 3 shows the results. A simple but optimized quality certification is predicted to achieve 94% of the optimal design's welfare. This design addresses the informational loss from misclassification, the multitasking moral hazard problem, and the aggregate underprovision of quality on average. It fails only at incentivizing heterogeneous production. However, as low willingness-to-pay consumers have a free and high-value outside option, the loss from eliminating variety at the bottom of the distribution of quality is small. Accordingly, the optimal cutoff for certification is slightly lower than the fourth cutoff of the best linear substitute.

VII.C.3 Competition: The new design changes the degree of vertical differentiation across firms and excludes a considerable fraction of competitors from the market. The semi-elasticity of substitution across firms decreases by 7.2% under the new design, weighted by enrollment.[47] Average markups grow considerably in response, particularly among previously low-rated firms that obtain high scores under the new system. A regression of markup changes among non-excluded plans (i.e., not in the lowest score group) on competition shows that an additional competing plan reduces the markup change by 0.3pp and an additional competing firm by 0.7pp. An additional competitor is also associated with a 1.8pp larger increase in quality following the regulation. However, while more competition stimulates quality provision, it also reduces the value of coarsening consumers' information. Conditional on the number of plans in a market—which partially accounts for the effect on vertical differentiation—an additional competitor reduces Spencian distortion by 5.4%, and the gains from coarsening information vanish at around 4.7 firms per market.[48] In equilibrium, welfare under full information is higher for only 9.9% of consumers.

VII.C.4 Regulatory preferences: CMS's preferences over equilibrium quality might include factors beyond consumers' surplus and firms' profits. For example, the most significant discrepancy between the best linear substitute design and CMSs is the relative weight placed on chronic condition management (Intermediate) relative to medical quality (Outcome). This difference could be due to a paternalistic view that consumers undervalue the future impact of worsening chronic conditions or because CMS is the residual payor for these future expenses. Thus, while the value of this attempt to shift the market is known only to

[47] Premium semi-elasticity, $(\partial D_j/\partial p_k)*(1/D_j)$, is used to avoid the effect of zero premiums on the analysis, as in Aizawa and Kim (2018).

[48] Markets with more plans are larger and more heterogenous in WTP. More plans allow firms to exert market power along consumers' WTP curve, which, on average, reduces welfare. Spencian distortion is measured as quality's absolute marginal welfare value under full information.

CMS, the cost of doing so can be computed. To do so, I compute the optimal certification of quality for a range of weights starting from the optimum and adjust the relative importance of the Intermediate and Outcome categories to span CMS's designs between 2009 and 2019.[49]

The results indicate that increasing the contribution of Intermediate relative to Outcome leads to a reallocation of investments from the second category to the first. However, the relative improvement in Intermediate quality rapidly plateaus as consumers' WTP for certified products deteriorates. Certification becomes less representative of the information consumers require to decide among plans; thus, its effect on enrollment choices decreases. This effect is compounded by an erosion of investment incentives since firms no longer benefit as much from certification, and welfare plummets together with quality and information. Overall, the results indicate that to justify the distortion observed in the data for 2015, CMS would have to value its small effect on chronic conditions by $14 billion per year. This loss in welfare is 17 times larger than the added investment it produces in the Intermediate category; Any subsidy that generates more than five cents in investment per dollar spent would outperform this attempt to skew quality with information. Scores are a poor nudging mechanism as it is inherently costly to steer consumers with information they do not value.

VII.C.5 Discussion: The results above have implications for scoring design beyond MA. The finding that optimal granularity is second order to optimal weighting indicates that the most salient feature of scores might be the least relevant one. Whether scoring by 5 letter grades, 1 to 5 stars with half-steps, or thumbs-up or -down, scores' performance might be determined primarily by their aggregation method. Moreover, this suggests that optimizing certifications might be better than disclosing more granular information in markets with moderate heterogeneity in WTP and good outside options (e.g., incremental technologies). This is bolstered by the finding that coarse scores approximate full-information outcomes.

These findings also imply a contradiction between efforts to regulate quality and disclose it. Neither quality nor consumers' information about it is monotonic in the ex-ante informativeness of scores.[50] More granular systems can lead to worse quality outcomes and exacerbate the effect of investment risk on quality variance. This is relevant for the joint efforts of CMS to promote and disclose quality (MedPAC, 2018) and likely also for many other markets where pay-for-performance and scoring policies coexist, such as in schooling,

[49] I use optimal certification for this exercise since it requires solving many optimal designs, and certifications are relatively faster to compute. Previous results established that certification approximates the welfare of the optimal design. See Online Appendix III for further details.

[50] Here, ex-ante means at the moment the designer chooses its scoring policy.

hospitals, and energy-efficient construction.

Finally, the design exercise reveals a method to address the multitasking moral hazard problem. These incentives to *game* the system have plagued various disclosure policies, sometimes leading to unintended consequences (Feng Lu, 2012; Reynaert and Sallee, 2021). The results show that firms' incentives can be aligned with regulatory objectives by designing aggregation weights properly. Moreover, they can allow heterogeneous firms to reach high-quality production through various paths, which enables an array of products that stricter minimum quality standard policies would not permit. Hence, if properly regulated, multidimensional quality can be beneficial rather than a source of harmful gaming. To induce meaningful total investment, however, some demand penalty must be imposed on those falling below a threshold. This finding contradicts recent advice given to Congress regarding eliminating "cliff effects" in insurer incentives in MA (MedPAC, 2020).

VIII Robust Scoring Design

In this final section, I consider scoring design under the *ignorance* assumption. In this case, consumers' preferences are only set identified, and the regulator cannot affect their quality beliefs. However, the regulator has already observed the impact of scores on enrollment and can leverage their assignment to marshal demand and induce quality investments. Knowing only that consumers' preferences are in some set Γ, the robust scoring design objective is to maximize total welfare under worst-case preferences:[51]

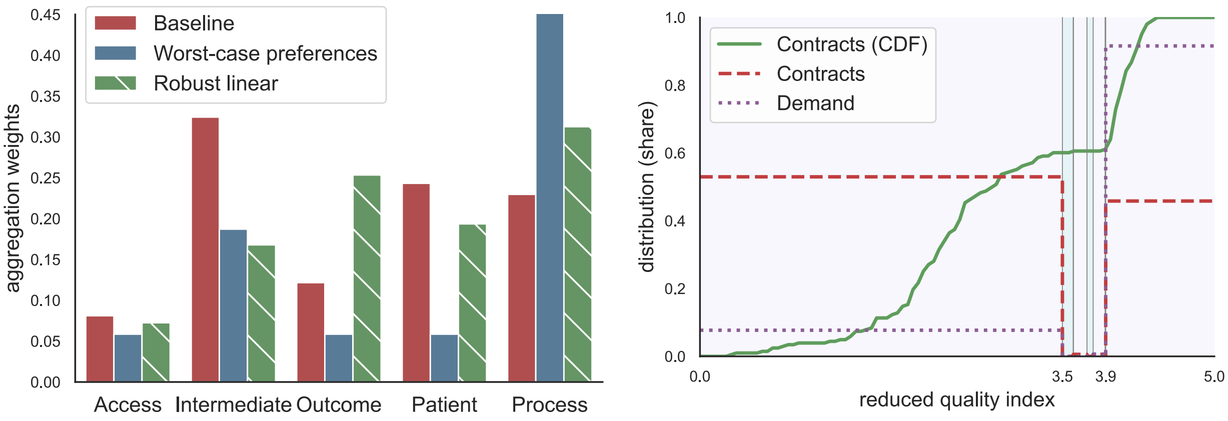
$$\max_{\psi\in\Psi}\min_{\gamma\in\Gamma}\int[\underbrace{CS(\psi,\boldsymbol{q})}_{\substack{\text{Consumer}\\ \text{surplus}}}+\rho^F\underbrace{\sum_f V_f(\psi,\boldsymbol{q})-I(\boldsymbol{x}^*_f(\psi),\mu_f)}_{\substack{\text{Insurer}\\ \text{profit}}}-\underbrace{\rho^G Gov(\psi,\boldsymbol{q})}_{\substack{\text{Government}\\ \text{spending}}}]dF(\boldsymbol{q}|\boldsymbol{x}^*(\psi)) \quad .$$

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. Importantly, this objective would remain the same if consumers had heterogeneous preferences for quality. The interior minimization is equivalent to a linear equilibrium constraint, which enables the use of the empirical design methodology.

Figure 7 shows the solution to this problem, and the first row of Panel B in Table 3 shows its worst-case welfare improvement.[52] The results reveal that surplus and profits could be

[51] Preferences γ are only relevant for consumer surplus because, conditional on identified fixed valuations for scores, the demand is independent of consumers' preferences for quality.

[52] To discipline the worst-case preferences, I restrict γ between half the lowest estimated value and twice the



(a) Aggregation weights

(b) Quality distribution and cutoffs

Figure 7: Robust certification design

Notes: Figure (a) shows the optimal robust and baseline aggregators and the worst-case preferences under the optimal weights. Preferences are not inversely proportional to weights due to non-uniform identified set bounds. Figure (b) shows the cutoff locations and contract distribution.

improved under these adverse regulatory conditions. The intuition behind these improvements is familiar. Scores shift demand away from underperforming products, which creates incentives for firms to invest. The new design assigns 1 star to all qualities at the bottom of the distribution and 5 stars to all high qualities, producing the pooling at the bottom effect. Since consumers' beliefs are fixed, the new design cannot address the misclassification problem but can still stir relative quality production and solve the multitasking moral hazard friction. However, since consumers' preferences are adversarial, the new weights create incentives for uniform quality production by being approximately inversely proportional to costs.[53] This makes attaining 5 stars more costly; it leaves only 45.8% of contracts in the top tier, which serve 91.5% of consumers, the remainder of products and consumers matching on 1 star. Finally, its granularity is that of a *padded* certification. While in expectation, all contracts obtain either 1 or 5 stars, the design allocates 5 intermediate levels in a narrow band to reduce insurers' investment risk and promote investment.

The robust design objective also serves as the worst-case scenario for those developed in the previous section. Panel B in Table 3 shows that in the worst-case scenario, total

highest among all quality dimensions. This means that the highest quality product can be worth anywhere between $4,133 and $44,984 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. This only affects welfare magnitudes, not design choices.

[53] 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.

welfare under the best linear substitute is positive, although to the detriment of consumers. Interestingly, it also shows that the worst case of the CMS-weighted optimal design (with reoptimized cutoffs) is better than that of the best linear substitute. CMS's weighting scheme is particularly well suited to address the robust problem and is nearly optimal within the class of linear aggregators. This observation suggests that CMS's design 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 earlier 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 and the previous section—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.

IX Conclusion

I study the problem of designing a scoring system for firms with market power over quality. Using detailed data from Medicare Advantage in 2009-2015, I show that scores shift demand across products and alter insurers' quality investments. Using variation in the scoring design, I specify and solve the problem of a welfare-maximizing regulator and find a constrained optimum. By decomposing and examining the solution, I derive empirical findings about scoring design in oligopolistic markets.

First, the results suggest that optimal designs involve coarsening consumers' information. Under full information, market power over quality leads firms to invest inefficiently. A coarse score can correct these incentives by shifting demand, creating penalties for underperforming firms. Second, I find that simple designs that are likely easy for consumers to interpret might also be remarkably effective. Hence, there is no inherent conflict between scoring for sophisticated or more behavioral consumers; they both react to scores, change their demand, and exert regulatory pressure on firms. Third, it can be extremely costly to use scores to steer quality production away from consumers' preferences. Skewing scores' informational content away from what consumers care about quickly erodes their informational value and regulatory power. Finally, both theoretical and empirical results show that transparency in scoring design is paramount for eliciting consumers' preferences and the score's effectiveness as an informational and regulatory policy.

My results support the growing theory on scoring design and point the way to several potential extensions. Incorporating market dynamics and measurement error would be helpful for scoring design in several markets with persistent investments and hard-to-measure outcomes. Accounting for data manipulation would help address challenges documented in nursing home scores and credit ratings. Finally, I assume that the quality domains and dimensions are fixed. How to define quality as a policy decision remains an open question.

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.

— and Gruber, J. (2011). Choice Inconsistencies among the Elderly: Evidence from Plan Choice in the Medicare Part D Program. *American Economic Review*, **101** (4), 1180–1210.

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.

Alé-Chilet, J. and Moshary, S. (2022). Beyond Consumer Switching: Supply Responses to Food Packaging and Advertising Regulations. *Marketing Science*, **41** (2), 243–270.

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

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.

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

Atal, J. P., Cuesta, J. I. and Sæthre, M. (2022). Quality regulation and competition: Evidence from pharmaceutical markets.

Aumann, R. J., Maschler, M. and Stearns, R. E. (1995). *Repeated games with incomplete information.* MIT press.

Ball, I. (2020). Scoring Strategic Agents. *Working paper.*

Barahona, N., Otero, C. and Otero, S. (2022). Equilibrium Eects of Food Labeling Policies. *Working paper.*

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

—, 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.

Blackwell, D. (1953). Equivalent comparisons of experiments. *The annals of mathematical statistics*, pp. 265–272.

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

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.

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

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.

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.

— 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.

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

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. 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.

Dai, W. D., Jin, G., Lee, J. and Luca, M. (2018). Aggregation of consumer ratings: an application to Yelp.com. *Quantitative Marketing and Economics*, **16** (3), 289–339.

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.

DECAROLIS, F., GUGLIELMO, A. and LUSCOMBE, C. (2020a). Open enrollment periods and plan choices. *Health Economics*, **29** (7), 733–747.

—, POLYAKOVA, M. and RYAN, S. P. (2020b). Subsidy design in privately provided social insurance: Lessons from medicare part d. *Journal of Political Economy*, **128** (5), 1712–1752.

DRAKE, C., RYAN, C. and DOWD, B. (2022). Sources of inertia in the individual health insurance market. *Journal of Public Economics*, **208**, 104622.

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

— 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.

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

FIORETTI, M. and WANG, H. (2019). How Does Risk Selection Respond to Quality Payments ? Evidence from Medicare Advantage. *Working paper.*

FLEITAS, S. (2020). Who benets when inertia is reduced? Competition, quality and returns to skill in health care markets. *Working paper*, p. 60.

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

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.

GLAZER, J. and MCGUIRE, T. G. (2006). Optimal quality reporting in markets for health plans. *Journal of Health Economics*, **25**, 295–310.

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., HENDEL, I. and WHINSTON, M. D. (2015). Equilibria in Health Exchanges: Adverse Selection versus Reclassification Risk. *Econometrica*, **83** (4), 1261–1313.

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

— and Kolstad, J. T. (2015). Health insurance for humans: Information frictions, plan choice, and consumer welfare. *American Economic Review*, **105** (8), 24492500.

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.*

— 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**, 24–52.

Hopenhayn, H. and Saeedi, M. (2019). Optimal Ratings and Market Outcomes. *NBER Working Paper Series*, pp. 1–39.

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

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.

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

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

— 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).

Kolotilin, A. and Zapechelnyuk, A. (2019). Persuasion meets delegation. *arXiv preprint arXiv:1902.02628.*

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

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. (2009). Measuring welfare losses from adverse selection and imperfect competition in privatized medicare.

Marone, V. R. and Sabety, A. (2022). When should there be vertical choice in health insurance markets? *American Economic Review*, **112** (1), 304–42.

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

MedPAC (2018). The Medicare Advantage Program: Status Report. *Report to Congress: Medicare Payment Policy*, pp. 287–306.

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

Miller, K. S., Petrin, A., Town, R. and Chernew, M. (2022). Optimal managed competition subsidies. *NBER Working Paper Series*.

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

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. *Working paper*, (April 2009), 1–45.

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.

Reynaert, M. and Sallee, J. M. (2021). Who benefits when firms game corrective policies? *American Economic Journal: Economic Policy*, **13** (1), 372–412.

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

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

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.

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

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.

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