For Online Publication: Online Appendix

Quality Disclosure and Regulation: Scoring Design in Medicare Advantage

Benjamin Vatter — September 19, 2022

Additional Tables

Table 1: Demand responses to scoring and model Robustness

(A) Stars and new MA enrollment			**(B) Weights and 5 stars coeff**		
2.5 stars	0.0212*	(0.011)	Access	0.262***	(0.055
3.0 stars	0.0450***	(0.011)	Intermediate	-0.339***	(0.068)
3.5 stars	0.0596***	(0.011)	Outcome	-0.958***	(0.203
4.0 stars	0.0823***	(0.011)	Patient	1.818***	(0.485)
4.5 stars	0.111***	(0.013)	Process	0.599***	(0.138)
5.0 stars	0.203***	(0.023)	**(C) Residual ranking test**		
N	55934	R^2 0.0516	Local rank	-0.00132	(0.00119)

Notes: Panel (A) displays the estimates of equation (2) in the main text. Panels (A) and (B) display robustness results discussed in Appendix I. (B) displays five different regressions matching equation 1, and (C) a test of whether consumers evaluate plans according to their star's local ranking, conditional on their score. Standard errors in parentheses are heteroskedastic robust. *p<0.05, **p<0.01, ***p<0.001.

Table 2: Plan quality response to design variation and robustness

	(I) - Main		**(II) - Uncensored**		**(III) - Quartiles**		**(IV) - Controls**	
Preexisting quality group (G_{lj})								
1 star			1.226***	(0.0669)				
2 stars	0.592***	(0.0601)	0.873***	(0.0603)			0.630***	(0.0630)
3 stars	0.163***	(0.0425)	0.393***	(0.0524)			0.185***	(0.0458)
4 stars			0.250***	(0.0469)				
2nd quartile					0.156***	(0.0387)		
N	195575		198189		194914		36464	
R^2	0.588		0.594		0.589		0.701	

Notes: This table presents the triple-difference estimates of quality responses to design changes. Column (I) shows the main estimates, as described in the main text. The remaining columns are associated to robustness checks discussed in Appendix I. Column (II) shows robustness to the domain censoring, column (III) to CMSs' cutoffs, and column (IV) to selection of controls. Standard errors in parentheses are clustered at the contract level. *p<0.05, **p<0.01, ***p<0.001.

Table 3: Demand estimation first stage

	Premium (p_{jmt})		Benefits (b_{jmt})	
Instruments				
Benchmark	0.00772***	(0.00180)	-0.00818***	(0.00118)
Rebate	0.398***	(0.00484)	-0.162***	(0.00299)
Other endogenous				
Benefits	1.289***	(0.00914)		
Premium			0.463***	(0.00301)
N	28830		28830	
R^2	0.887		0.896	

Notes: This table reports the first stage estimates of the second step of the demand estimation. All regressions include controls for other observable product characteristics and contract-year fixed-effects. Standard errors in parentheses are heteroskedasticity robust. *p<0.05, **p<0.01, ***p<0.001.

Table 4: Remaining estimated demand coefficients

	coefficient	std. err		coefficient	std. err
TM x Demographics (λ^d)			**Switch: Across insurers**		
ESRD	1.284***	(0.277)	Health - Excellent	-3.104***	(0.153)
Attended college	0.037	(0.044)	Health - Very Good	-3.138***	(0.110)
College degree or higher	0.141**	(0.045)	Health - Good	-3.158***	(0.111)
Disabled	-0.040	(0.104)	Health - Fair	-3.168***	(0.151)
Employer-Sponsored	1.017***	(0.041)	Health - Poor	-2.835***	(0.272)
Female	-0.076	(0.067)	**Switch: Within insurer**		
Graduated high school	0.015	(0.052)	Health - Excellent	-1.948***	(0.223)
Medium Income	-0.067	(0.081)	Health - Very Good	-2.271***	(0.144)
High Income	0.096	(0.086)	Health - Good	-1.910***	(0.155)
Asian indicator	0.176	(0.117)	Health - Fair	-1.848***	(0.221)
Black indicator	-0.090	(0.062)	Health - Poor	-1.449***	(0.368)
Hispanic indicator	-0.341***	(0.066)			
Switch: MA to TM			**Plan Attributes**		
Health - Excellent	-3.727***	(0.113)	Enhanced drug benefits	0.093***	(0.012)
Health - Very Good	-3.930***	(0.083)	Dental exam	-2.409***	(0.083)
Health - Good	-3.569***	(0.082)	Dental fluoride	-0.438***	(0.024)
Health - Fair	-3.717***	(0.111)	Dental Xray	0.682***	(0.040)
Health - Poor	-3.796***	(0.208)	Hearing fitting	-0.106***	(0.025)
Enrollment choices	36447		Plan-county-year	29004	

Notes: This table displays estimates for the demand coefficients not presented in Table 1 in the main text. Unadjusted heteroskedastic standard errors are in parentheses. *p<0.05, **p<0.01, ***p<40.001.

Table 5: Data descriptive statistics

	TM		MA			MA	
MCBS - Enrollees					**CMS - Plans**		
Female	0.527	(0.499)	0.555	(0.497)	Bid	9439.215	(1058.175)
Age	72.257	(10.234)	73.025	(8.994)	Benchmark	10146.541	(964.674)
Part B premium	103.003	(41.743)	111.724	(20.140)	Benefits	885.201	(478.974)
Income	49.454	(63.854)	44.417	(57.179)	MA Premium	351.918	(493.690)
ESRD	0.007	(0.082)	0.003	(0.057)	Rebate	765.506	(602.134)
Disabled	0.113	(0.316)	0.098	(0.298)	Part D premium	285.546	(291.295)
Health - Excellent	0.183	(0.387)	0.176	(0.381)	Market share	0.024	(0.035)
Health - Very good	0.306	(0.461)	0.327	(0.469)			
Health - Good	0.284	(0.451)	0.290	(0.454)			
Health - Fair	0.144	(0.351)	0.148	(0.355)			
Health- Poor	0.060	(0.238)	0.047	(0.212)			
Observations	58211		12416			139323	

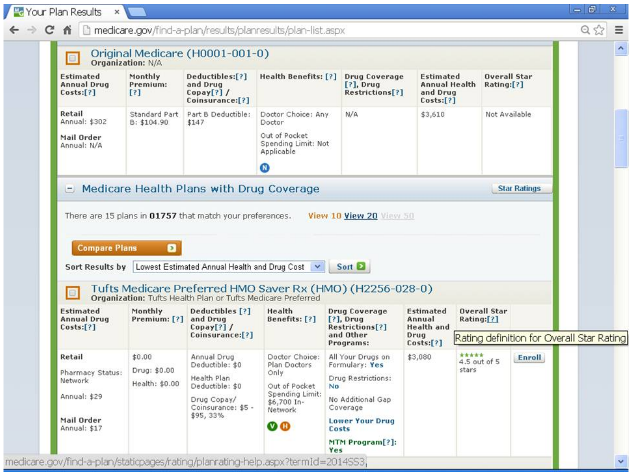
Notes: This table presents the data means and standard deviations (in parentheses). Observations in the MCBS enrollment data are weighted by nationally representative multipliers. All values are Health-CPI adjusted to 2015 values. Plan attributes are yearly.

Appendix I: Institutional Details and Descriptive Evidence

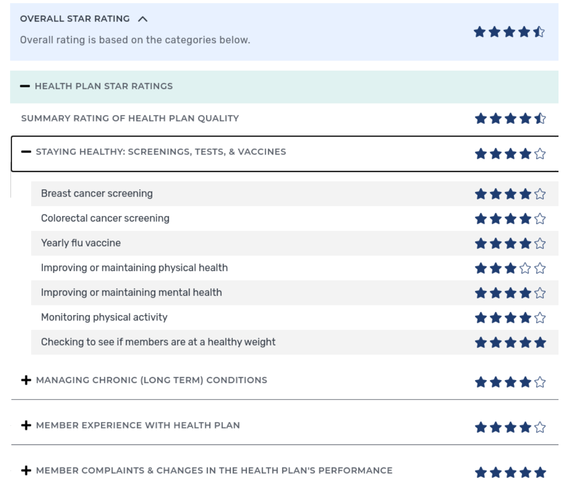
Enrollment platform: MA organizes plan offerings in a unified shopping platform and regulates contract characteristics. Figure 1a displays the view of the platform in 2015. The platform presented consumers with the Original Medicare option at the top and MA plans at the bottom. Each plan displayed its estimated deductibles, cost-sharing rules, and monthly premiums. The system showed the "Estimated annual health and drug costs" closely related to the benefit levels used in the main text. Also, the system included the MA Star Rating for the plan next to the enrollment button. Clicking on the question mark in the column name revealed the basic construction details of this rating, as in Figure 1b.

Pricing regulation: Price and coverage regulation in MA operates through a process known as bidding. Every year, insurers submit insurance plan offerings, listing each plan's participating counties, cost-sharing attributes, actuary-certified estimates of expected expenses, plan-level estimates of administrative costs, and profit margins. The bidding procedure combines these data to form two components: First, the revenue required to cover expenses and margins related to standard and mandatory Medicare coverage, which CMS calls the *bid*. In the main text, I call this value the plan's *price*. Second, the revenue required to cover supplementary benefits, such as lower copays and maximum out-of-pocket amounts, which I refer to as the plan's additional *benefits*. These additional benefits are not optional and exclude dental, vision, hearing, or Part D prescription coverage.

The bidding process compares a plan's bid against a benchmark related to TM's cost. CMS computes plan benchmarks by averaging TM's fee-for-service costs in every county the



(a) Plan Compare website



(b) Plan Compare rating details

Figure 1: Medicare plan finder view

Notes: Figure (a) shows a view of the Medicare Plan Finder platform from 2015. Figure (b) shows the detail presented to consumers after clicking on the details button in 2022. A similar view was offered in 2015.

plan operates in, using weights proportional to the plan's expected enrollment. CMS pays plans bidding above the benchmark an amount equal to the benchmark per enrollee, and enrollees pay the difference in what is known as the *basic* MA premium. For plans bidding below the benchmark, CMS pays their bid plus a rebate equal to a fraction of the difference.

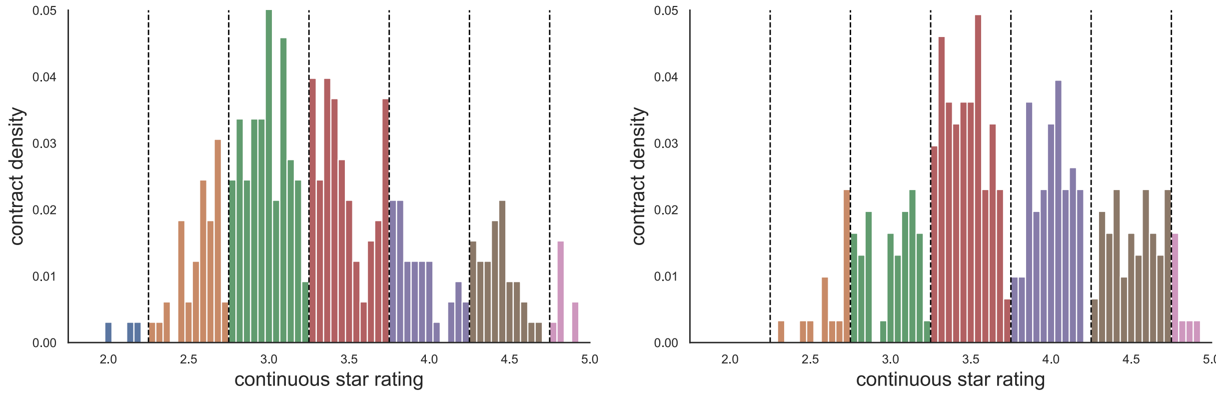
Additional benefits allow MA insurers to offer more generous coverage than TM. However, to provide these benefits, insurers must fund them through either premiums or rebates. Specifically, insurers must use every dollar of rebates to either fund benefits or buy down non-MA enrollee premiums. The latter includes the part B premium CMS charges to every enrollee regardless of their choice between TM and MA and any part D prescription drug premium the plan might charge. Any additional benefits not funded by rebates are paid directly by the consumer under a *supplementary* MA premium.

Overall, the following equations summarize this regulation.

$$\begin{aligned}\text{Rebate}_j &= \rho_j \max\{B_j - p_j\}\\ \text{Premium}_j &= \underbrace{\max\{p_j - B_j, 0\}}_{\text{basic}} + \underbrace{\max\{b_j - \text{Rebate}_j, 0\}}_{\text{supplementary}}\\ \text{Payment}_j &= \min\{p_j, B_j\} + \text{Rebate}_j + \text{Premium}_j\end{aligned}$$

Where b_j is the plan's additional cost-sharing benefits, measured in cost-savings for the average unit-risk consumer, and ρ_j is the rebate share.[1] Put in perspective, per member per month, the average MA plan in 2015 (by enrollment) submitted a price of \$700, additional

[1]The rebate share has varied over the years and, since 2012, depends on plans' rating in previous years.



(a) Year 2012

(b) Year 2015

Figure 2: Distribution of underlying continuous aggregate star rating

Notes: The figures show the histogram of aggregate contract quality before rounding. Every contract within a same-color bin obtains the same rating, denoted on the horizontal axis.

benefits equal to \$70, and faced a benchmark of \$782. Among plans with a non-zero premium (43.8%), the average was \$73.5, with 13.4% coming from the basic MA premium. More than half of enrollees chose a zero-premium plan (58%). However, 83.9% of plans had a zero basic premium, resulting in a rebate averaging \$63.8. Every MA plan offered additional consumer benefits, averaging an actuarial value for medical services of 87.2%.[2]

Comparison to previous descriptions of the market: There is some discrepancy in the literature regarding the bidding process. A recent release of information by CMS, containing the complete bidding data, software, and instructions sent to insurers for 2009-2015, informed the description above and solved some of these discrepancies. For example, bidding is often misdescribed as taking place at the plan-county level. However, insurers can segment their contracts across counties and submit separate bids for each segment. The fact that 99% of all plans offered in more than one county chose not to segment suggests the gains from bidding independently across counties are small, and so are the losses from treating competition as occurring at the county level.

Another distinction is that premiums are often described as the positive difference between bids and benchmarks. However, this is only the definition of the *basic* MA premium, which constitutes but 12% of the total MA premium paid by consumers. The standard model also describes rebates as the positive difference between benchmarks and bids. Thus, each plan should only have either a premium or a rebate. The data refute this model as 45% of plans have both premium and rebate.[3]

[2]Actuarial values of MA are computed using public CMS software.

[3]Using the wrong model can have implications for previous work. In particular, previous work inferred bids from premiums or rebates. Depending on the scenario, these inversions can lead to erroneous bids and linkages between prices and subsidies.

The Star Rating Program: Improving the quality of care and beneficiary general health is one of CMS' most important strategic goals (Centers for Medicare and Medicaid Services, 2016). To this end, CMS has undertaken several initiatives to gather and display the quality of MA plans. Following the Balanced Budget Act of 1997, CMS began collecting information on multiple quality measures through surveys and insurer reports. A summary of the gathered data was first presented to consumers in the November 1999 edition of *Medicare & You*, a handbook mailed annually to Medicare-eligible enrollees. The impact of this first implementation was noticeable, as studied by Dafny and Dranove (2008). In 2007, CMS began summarizing the quality information into five quality domains (e.g., "Helping You Stay Healthy"), with values described by one to five stars. In 2009, the star rating program took on its current form, with a single overall rating displayed to consumers next to the plan's name, premium, and cost-sharing attribute. The Supplementary Material presents the formulas used to construct plan scores, adjustment factors, and the adjustment to rebate shares that depend on previous scores.

Data: Details about the data construction are provided in the Supplementary Material. Table 5 displays descriptive statistics of the combined plan-county-level public data and the administrative MCBS individual-level data.

Understanding the design: The Star Ratings' design consists of fundamentally two components: the overall contribution of each category to the score and the cutoffs used to discretize the measures underlying each category. Category contributions are easily observed as measure weights are uniform within a category: 3 for Outcome and Intermediate Outcomes, 1.5 for Patient and Access, and 1 for Process.[4] Therefore, it suffices to know the number of measures in each category, which is visible on the enrollment platform. CMS informs consumers of the scores and what they capture in a yearly booklet sent to every beneficiary called *Medicare & You* and on their website. Additionally, large changes to the design are often covered in news articles and discussed in promotional material by insurers.

Measure-level cutoffs are less visible, but two features facilitate their understanding. First, their variation over time is minimal, which allows consumers to learn them. As Medicare beneficiaries are either retirees or disabled, they have likely had significant interaction with their local healthcare markets. They would, therefore, understand the distribution of quality inputs for insurers. By looking at the relative distribution of ratings in their markets, consumers can use their experience to form a partial understanding of cutoffs. For example, if only a small fraction of plans have five stars, but consumers know that there are many high-quality providers and the contribution of the Outcome category is large, it stands to reason that the average five-star cutoff within the Outcome category is very high.

The second facilitating feature is that, as I discuss in Appendix III, cutoffs need not be understood individually. As cutoffs within a category are averaged and then discretized in the overall score, they are well approximated by cutoffs to aggregate quality. These cutoffs

[4] The only exception are newly introduced measures, which always get a weight of one.

are fewer and benefit from the same stability as their measure-level sources.

Testing consumers' understanding: The ideas expressed above are partially testable. In particular, we can test the null that consumers are ignorant of scoring weights by examining the correlation between consumers' preferences for scores and their underlying design. If consumers are unaware of changes to the underlying scoring design, changes to category weights should not affect their choices, conditional on the scores assigned to each plan. Figure 3b in the main text suggests that this is not the case, but it can be tested further by adapting the regression of equation (2) in the main text to be:

$$y_{ijt} = \alpha_{r(jt)} + \beta_{\tilde{r}k} w_{kt} \mathbb{1}\left\{r(jt) = \tilde{r}\right\} + \boldsymbol{x}_{jt}\boldsymbol{\lambda} + \mu_{m(i)} + \epsilon_{ijt}, \tag{1}$$

where w_{kt} is the weight of some category k in year t and $\tilde{r}$ is an arbitrary rating level. Under the conjecture that consumers are ignorant of changes to the scoring design, $\beta_{\tilde{r}k}$ should be zero for all $\tilde{r}$ and k. Panel B in Table 1 shows the estimated $\beta_{\tilde{r}k}$ for every category k and $\tilde{r} = 5$, rejecting the null of ignorance at 1% confidence.

Scores as rankings: Consumers who are imperfectly informed about scoring design might rely on the distribution of scores in their market to make additional quality inferences. In particular, consumers might use scores to form a ranking of their options, giving scores an ordinal interpretation. To evaluate this, I test whether the residual of individual demand in equation (2) of the main text systematically correlates with the local rank of a plan within its county. Panel C of Table 1 shows the estimated coefficient of the rank, where 1 indicates the highest scoring plan in the county (ties are assigned equal rank). A firm fixed effect is added to the residual regression to partially address selective entry by firms to counties with fewer competitors.[5] The estimated coefficient is negative, indicating that a higher rank plan is preferred. However, the effect is statistically insignificant, and its magnitude is negligible.

Imperfect quality control: MA features substantial cross-sectional variation in quality. Before their final discretization, the distribution of continuous ratings is well dispersed, as shown in Figure 2.[6] While it might appear contradictory for a firm to invest in quality in the interior of a rating interval—as consumers do not observe it—, there is a simple explanation for it. Investments in MA are contractual arrangements with providers and third-party services. Insurers can change their quality by restructuring their network and forming incentive contracts, but the final delivery of quality is rarely in their control. For example, an insurer can expand its physician network to reduce waiting times for primary care appointments. However, a harsh flu season can increase the burden on physicians and result in higher wait times than expected by the insurer when optimizing the network.

Robustness of quality responses to scoring design: The analysis done in section IV.B

[5] Removing firm fixed effects slightly increases the magnitude but does not change the result qualitatively.

[6] Star ratings of 1 and 1.5 are rarely observed. Rating between 2 and 3 are less rare but not often provided by top insurers. Otherwise, large firms provide contracts covering the range of stars.

in the main text suggests a causal effect of scoring design on quality. Column I in Table 2 presents the estimates matching equation (3) in the main text. This analysis censors the quality domain, dropping plans whose preexisting quality falls within the first or last score. This avoids inflating the results due to reversion to the mean. Column II shows the effect of removing this censoring which, as expected, increases the effects. The analysis also relies on CMS's definition of the cutoffs, which might be subject to influence by insurers due to lobbying. Column III in the table shows that the results hold qualitatively if one compares the second to the third quartiles of preexisting quality rather than relying on cutoffs.

Recent research has raised concerns about staggered difference-in-differences designs similar to the one used in this analysis (Goodman-Bacon, 2021; Callaway and Sant'Anna, 2020; Baker *et al.*, 2021). However, the structure of the treatments used in this work differs from the canonical example used in this literature. Unlike standard staggered differences-in-differences, where the same treatment is assigned to different units over time, measure entry can be seen as different treatments assigned to different units. However, regardless of this distinction, the concern remains that by aggregating the effect of different treatments, the pooled regression used in the main analysis might deliver a biased estimate of the treatment effect. To alleviate this concern, I structure the data behind this analysis as a *stacked regression estimator* (Baker *et al.*, 2021). In this structure, the data is propagated so that each event and unit are directly matched with all their controls. Column IV in Table 2 shows robustness to randomly selecting 20% of controls rather than using all available observations. Additionally, Table 6 shows the estimated coefficient on the main analysis separately for each measure. The coefficient aggregates the effect on plans in groups 2 and 3 relative to 4 to reduce the number of estimates. The results show that all effects are positive, and most are significantly so. Therefore, the results are unlikely to be driven by a negative or non-convex weighting of the underlying individual events.

Table 6: Plan quality response to design variation - Robustness

Measure	coefficient	std. err	Measure	coefficient	std. err
Access & performance	0.858***	(0.213)	MTM program completion	0.123	(0.187)
BMI assessment	0.488***	(0.131)	Medication reconciliation	0.772***	(0.145)
Breast cancer screenings	0.133*	(0.054)	Members leaving plan	0.256*	(0.125)
Enrollment timeliness	0.426***	(0.117)	SNP management	0.266	(0.272)
Improving bladder control	0.240*	(0.094)			

Notes: This table presents the coefficient estimated by measure in the triple-differences regression used to evaluate quality responses to design changes. The coefficient pools the response of plans with preexisting quality falling in 2 or 3 stars, relative to those of 4 stars. Standard errors in parentheses are clustered at the contract level. *p<0.05, **p<0.01, ***p<0.001.

Appendix II: Model, Identification, and Estimation

Consumers' total premium: The total premium consumers pay in the model (p_{jmt}^{total}) is a sum of their mandatory Part B premium (p_i^B), the MA plan's premium (p_{jmt}^C), and a Part D premium (p_{jmt}^D) if the plan bundles prescription drug coverage. The result is $p_{ijmt}^{\text{total}} = p_i^B - PR_{jmt}^B + p_{jmt}^C + p_{jmt}^D - PR_{jmt}^D$ where PR_{jmt}^B and PR_{jmt}^D are rebate dollars allocated by the insurer to reduce consumers' part B and part D premiums.

Insurers' added revenue and rebate allocation: CMS requires plans to allocate their rebates among benefits or premium reductions (see Appendix I). Each plan determines a fraction of rebates to allocate to Part B premium reductions (κ_{jmt}^b), Part D reductions (κ_{jmt}^d), extra benefits (κ_{jmt}^e), and increasing consumers' coverage on standard Medicare-covered health care services ($1 - \sum_{l \in \{e,d,b\}} \kappa_{jmt}^l$). Because premium reductions are payments directly to CMS or a transfer within the firm, they do not add to the firm's revenue. CMS strictly regulates improvement to standard Medicare coverage, requiring insurers to submit cost assessments based on CMS utilization models certified by actuaries. Because of this, I assume that plans offer these additional benefits at cost. The only remaining free source of rebate revenue is the κ_{jmt}^e. Therefore, the additional revenue of plan j in market m at year t, $R(p_{jmt}, \boldsymbol{z}_{jt})$ is given by the sum of its Part D premium (p_{jmt}^D) and any rebates allocated to additional benefits ($Rebate_{jmt}(p_{jmt})\kappa_{jmt}^e$). I assume that in the short run, each plan's rebate allocation fraction is an exogenous feature.[7]

Premiums and benefits computation: As described above, consumers' premiums are the sum of Part B, C, and D premiums, minus reductions. Part B and D premiums are treated as exogenous attributes. The former is an individual-specific element, and the bundled prescription drug coverage plan determines the latter. The Part C premium is the sum of the basic and supplementary premiums. The basic premium is equal to the positive difference between bid and benchmark, ($\max\{p_{jmt} - B_{jt}, 0\}$). The supplementary premium equals the fraction of additional Medicare cost-sharing plan benefits ($\bar{b}_{jmt}$) not financed by rebates, $\bar{b}_{jmt} - Rebate_{jmt}(p_{jmt})(1 - \sum_{l \in \{e,d,b\}} \kappa_{jmt}^l)$. Reductions are determined according to rebate allocation rules described above.

The total benefits of a plan correspond to the sum of its mandatory TM benefits (b_0), additional Medicare cost-sharing benefits ($\bar{b}_{jmt}$), and extra benefits ($\kappa_{jmt}^e Rebate_{jmt}(p_{jmt})$). The first two types of benefits are treated as exogenous. The first is a regulatory level dictated by CMS, and the second is a product attribute. The extra benefit term varies with the rebate and depends on the plan's bid. Hence, both premiums and benefits are endogenous components of the model, albeit to different degrees.

Model flexibility in quality provision: The inefficiency in quality provision found in the

[7]For the 13% of plans without rebates in the data, I assume that all counterfactual rebates would go to cost-sharing standard benefits. This assumption's effect is minimal on consumer premiums and firm revenue and does not affect this stage's estimates.

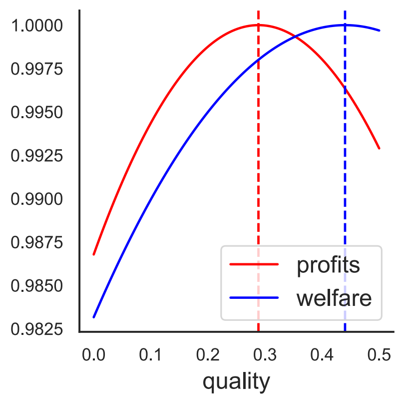
main analysis plays an important role in the resulting scores. I consider a simplified version to show that the model is flexible enough to accommodate both over and underprovision of quality. I ignore subsidies, investment risks, multidimensional quality, and multiproduct firm incentives to highlight the mechanisms. Consumers in the simplified model choose a plan to maximize an indirect utility $u_{ij} = \alpha_i p_j + \gamma \mathcal{E}[q|r_j, \psi] + \xi_j + \epsilon_{ij}$, where all terms are as in the main analysis. The utility of their outside option is normalized to zero, up to a logit error. Ignoring investment risk and focusing on single product firms, the insurer owning plan j chooses quality and prices to maximize $\pi_j(\boldsymbol{q}, \boldsymbol{p}, \psi) = D_j(\boldsymbol{q}, \boldsymbol{p}, \psi)(p_j - \theta' q_j - c_j) - \mu_j q_j^2$.

In this simplified model, under a scenario of full information and a monopolist firm, it is easy to find conditions under which quality is over or underprovided. In particular, by applying the logic of Spence (1975) and comparing the first-order conditions of the monopolist and the regulator, it is straightforward to show that quality will be overprovided if $\frac{1}{D} \int D_i \frac{\gamma}{\alpha_i} dF(\alpha_i) > \frac{\partial p}{\partial q}$, efficiently provided if this condition holds with equality, and underprovided if the inequality is reversed. On the left, $D_i \frac{\gamma}{\alpha_i}$ is consumer i's valuation for a marginal increase in quality, measured in units of premiums and weighted by her choice probability. Thus, the left-hand side of this equation is the weighted average valuation for marginal quality increases. On the right, we have the increase in price associated with a marginal increase in quality. This result is a direct analog to Spence's first proposition.

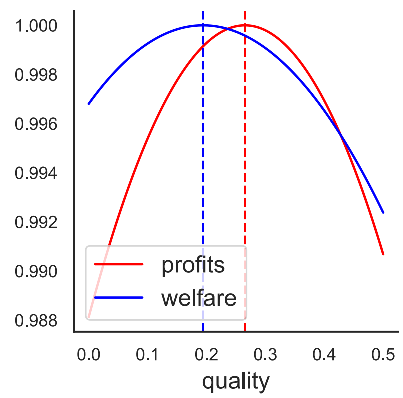
To illustrate that this condition can lead to different efficiencies of quality production, Figures 3a and 3b show simulations for two markets that differ only in their distribution of price preferences but generate opposing quality outcomes. Figure 3c shows that this extends to markets with more than one firm. In particular, it presents simulations for a duopoly market where consumers preferences are $\alpha_i \in \{1, 1.5, 3\}$, distributed with probability $[(1-x)/2, x, (1-x)/2]$, and $\gamma = 0.5$. Consumers have fixed unobserved preferences for firm 1 given by $\xi_1 = 2$, while their preference for the second firm, ξ_2, varies in the simulation. Firms have equal cost functions ($\theta = 0.2, \mu = 0.1, c = 0$). In the simulations, firm 1 always underprovides quality relative to the social optimum, but firm 2 might over or underprovide depending on consumers' preferences. In particular, it overprovides quality when consumers' WTP for quality is more heterogeneous and when ξ_2 is smaller.

Adapting the Star Ratings' design to the category-level: CMS' design involves hundreds of measures and cutoffs that vary over time. Using it fully would require identifying consumers' beliefs and insurers' investment costs for each quality measure. This task is both untractable and likely not an accurate representation of reality. For example, it seems implausible that an insurer can invest in improving the rate at which physicians review their patients' medication while decreasing the rate at which the same physicians assess their patients' pain. CMS's quality categories capture this correlation and dependence across measures and form a natural lower-dimensional space to evaluate quality.

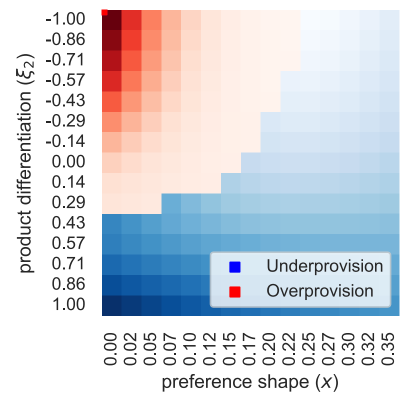
The baseline design is, intuitively, easy to approximate at the average-category quality level. The star rating of a contract is the weighted sum of many step functions, rounded to



(a) Underprovision



(b) Overprovision



(c) Duopoly

Figure 3: Efficiency with heterogeneous price preferences

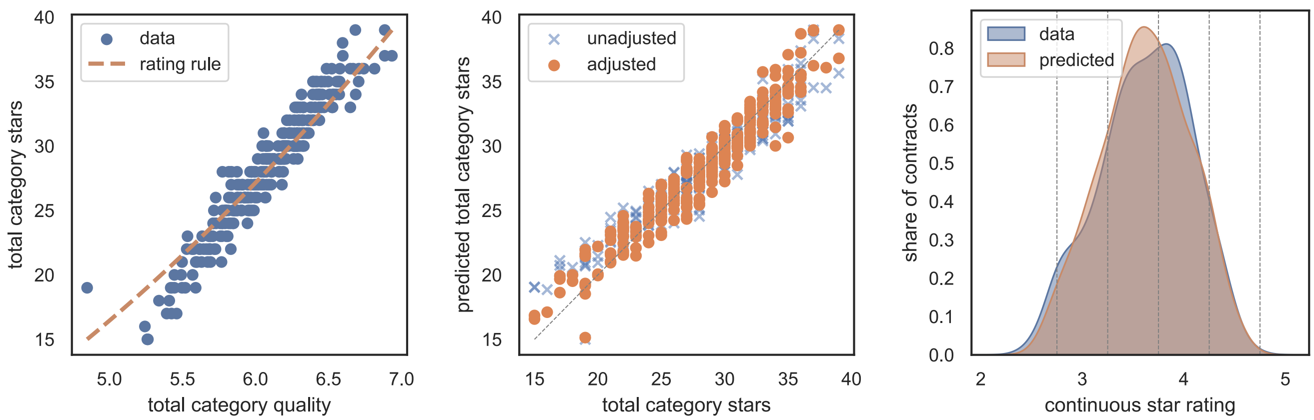
Notes: Figures (a) and (b) display scenarios of under and overprovision in monopolistic markets. Their common configuration is $\gamma = 0.5, \xi = 2.0$, $\theta = 0.2$, $\mu = 0.1$, $c = 0$. Price preferences take on values in $\{1.0, 1.5, 3.0\}$. In (a) the probability of each point is $[0.2, 0.6, 0.2]$, and in (b) $[0.4, 0.2, 0.4]$. Welfare and profits have been normalized to have a maximum of 1 for display only. Figure (c) illustrates the efficiency of quality in a duopolistic market. Red blocks indicate underprovision by the second firm and blue blocks overprovision. Lighter color tones indicate lower welfare losses from quality distortions.

the nearest half-star. Therefore, a smooth function could approximate each step function and have most of its fitting error rounded out. As the weights are category-specific and the distribution of quality within a contract is similar across firms, we could use a single smooth function per category to map average category-level investment to total category score.

The approximation procedure follows this intuition: First, for each category year, I fit a bounded polynomial that maps the sum of each plan's qualities in a category onto the total sum of measure-level stars; Second, I capture minor systematic differences across firms by regressing the approximation error of step 1 on indicators for plan-type, state, firm, and year. I also include the number of measures in the category and rating adjustment factors. The predicted value is added as a dispersion adjustment; Finally, I adjust category weights to maximize the fit, solving a least-squares problem subject to the constraint of positive weights. Figure 4 illustrates the three steps involved in this procedure.

Comparing the predictions of this procedure with the data for each year, the R^2 ranges between 0.91 and 0.946. The maximum absolute error of the model is only half a star, and 78% of plans get the same star rating as in the data. This remaining error is added to the adjustment factor such that the model predictions are exact in the baseline.

This approximation rule is only used three times in the paper. First, when estimating investment costs. The approximation error introduced there is relatively small as the rule is exact in the baseline, and only marginal changes are considered. Second, when estimating consumers' beliefs and preferences for quality under the informed choice assumption. This approximation is used to compute the quality domain that can achieve each rating in the baseline. It is mostly harmless and can be thought of as obtaining the quality partition



(a) Step 1: Intermediate 2015 (b) Step 2: Intermediate 2015 (c) Step 3: Overall 2015 rating

Figure 4: Scoring rule approximation

Note: These figures illustrate the scoring rule approximation used to reduce the complexity of the problem.

associated with each score. Finally, I use it to simulate the baseline in the counterfactual analysis. This is where the approximation is truly leveraged, as I compute scores under the baseline rule for counterfactual quality outcomes.

Star Ratings as monotone partitional scores: The Star Ratings are the result of a weighted average of monotonic measure-level step functions, rounded to the nearest half. As the class of monotone partitional scores is closed under addition and positive scalar multiplication, the weighted sum of measure-level scores is monotone partitional. It is also easy to verify that rounding does not break the monotonicity of the scores. Hence, the MA Star Ratings are monotone partitional before any approximation to average category quality.

At the category level, the above approximation suggests that we can view the Star Ratings as the rounded weighted sum of five monotonic continuous functions, each taking average category-level quality as an argument. The sum of monotonic functions on different domains is strictly monotonic on the product domain, and the rounding maps the values to finitely-many signals. Therefore, the Star Ratings at the category level are monotone partitional.

Demand identification - endogeneity of scores: Quality—as price and benefits—is correlated with consumers' unobserved preferences for plans. However, their discretization into scores limits the effect of this endogeneity on demand estimation. To validate this hypothesis, I rerun the second stage demand estimates, explicitly separating $\eta_{c(j)t}$ into a star-year fixed effects (FE) and contract FE. I consider the endogeneity of the star-year FE and treat it using two sets of instruments. The first is each plan's quality in a category relative to each other category, including every pair-wise combination. As noted in the scoring design section, firms' incentives to allocate investments across categories ignore consumers' preferences for plans or quality. They are driven exclusively by investment costs and their relative reward in the scores. Second, I consider the relative weight given to each pair of categories by the design in each year. I also include the interaction between these two

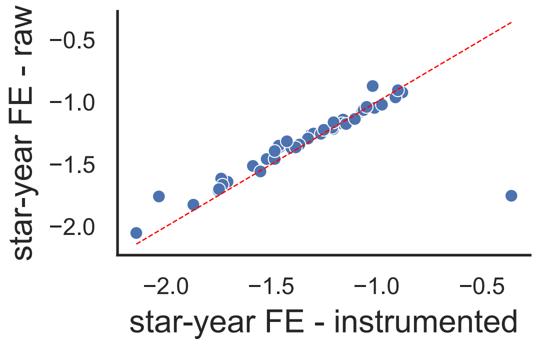


Figure 5: Estimates of consumers' score valuation with and without instruments
Note: This figure illustrates the star-year FE estimated in the second stage of the demand estimation. Each point represents the value of a given star-year, with its vertical position indicating its value without instruments and its horizontal position with instruments. The red dashed line marks the equality of estimates.

instruments. Intuitively, firms allocate a larger share of their investments to the categories in which they are more efficient. When the design changes and increases the relative reward of investing in one category relative to another, firms that have a relative cost advantage in this category will obtain higher scores. Figure 5 shows that the estimated FEs with or without the scores are nearly identical, differing on average by less than 3%. Intuitively, the instruments capture variation stemming only from design and costs, which are highly predictive of score changes within a contract. Therefore, the analysis of the main text does not use these instruments to identify consumers' WTP for star years. Instead, it uses design variation to disentangle beliefs from preferences for quality, given consumers' WTP for scores.

Demand identification - proof of Proposition 1: I begin by proving the result for the case of informed choice under scalar quality. Throughout, I assume the identification of the score-year fixed effect $\eta_{rt} = \gamma' \mathcal{E}[q|\psi(q) = r] + \bar{\eta}$, up to an additive constant $\bar{\eta}$, which in the context of MA, corresponds to the mean valuation for MA relative to TM. The proof depends on the following preliminary lemma.

Lemma 1. *Let f, g be two distinct, continuous, strictly positive densities, supported on $[0,1]$. Then, there exists $\underline{x} < \tilde{x} < \bar{x} \in [0,1]$ such that either $\mathbb{E}_f[x|x \in (\underline{x}, \tilde{x})] \geq \mathbb{E}_g[x|x \in (\underline{x}, \tilde{x})]$ and $\mathbb{E}_f[x|x \in (\tilde{x}, \bar{x})] \leq \mathbb{E}_g[x|x \in (\tilde{x}, \bar{x})]$ with one of the inequalities strict, or the analogous statement hold with the roles of f, g reversed. Also, there exists another $\underline{x} < \bar{x} \in [0,1]$ such that $\mathbb{E}_f[x|x \in (\underline{x}, \bar{x})] = \mathbb{E}_g[x|x \in (\underline{x}, \bar{x})]$*

Proof. As f and g are distinct and have common support, they must cross at an interior point $\tilde{x} \in (0,1)$. By continuity $\exists \epsilon > 0$ such that, without loss of generality, $f(x) > g(x)$ $\forall x \in (\tilde{x}, \tilde{x} + \epsilon)$ and with the weak opposite inequality in $(\tilde{x} - \epsilon, \tilde{x})$. Define $h_f(x, \epsilon) = f(x)/(F(\tilde{x} + \epsilon) - F(\tilde{x}))$ and analogously for g, where F is the cumulative of f and G that of g. Note that $\forall \tilde{\epsilon} \in (0, \epsilon)$ we have that $h_f(\tilde{x}, \tilde{\epsilon}) < h_g(\tilde{x}, \tilde{\epsilon})$, and that both $h_f(\cdot, \tilde{\epsilon})$ and $h_g(\cdot, \tilde{\epsilon})$ are continuous densities integrating to one within $(0, \tilde{\epsilon})$ and therefore intersect at an interior point. Pick $\bar{\epsilon} \in (0, \epsilon)$ such that $h_f(\cdot, \bar{\epsilon})$ and $h_g(\cdot, \bar{\epsilon})$ intersect only once at a point $\hat{x}$. Denote $\bar{x} = \tilde{x} + \bar{\epsilon}$. Then we have that $\mathbb{E}_f[x|x \in (\tilde{x}, \bar{x})] - \mathbb{E}_g[x|x \in (\tilde{x}, \bar{x})] =$

$$\int_{\tilde{x}}^{\bar{x}} (h_f(v, \bar{\epsilon}) - h_g(v', \bar{\epsilon}))vdv = \int_{\tilde{x}}^{\hat{x}} \underbrace{(h_f(v, \bar{\epsilon}) - h_g(v', \bar{\epsilon}))}_{<0} vdv + \int_{\hat{x}}^{\bar{x}} \underbrace{(h_f(v, \bar{\epsilon}) - h_g(v', \bar{\epsilon}))}_{>0} vdv$$
$$> \hat{x} \int_{\tilde{x}}^{\hat{x}} (h_f(v, \bar{\epsilon}) - h_g(v', \bar{\epsilon}))dv + \hat{x} \int_{\hat{x}}^{\bar{x}} (h_f(v, \bar{\epsilon}) - h_g(v', \bar{\epsilon}))dv = 0$$

This proves the first inequality. The proof for the second is analogous applied to $(\tilde{x} - \epsilon, \tilde{x})$. The equality part follows from defining $w(\lambda) = \mathbb{E}_f[x|x \in (\underline{x}, \tilde{x} + \lambda\bar{\epsilon})] - \mathbb{E}_g[x|x \in (\underline{x}, \tilde{x} + \lambda\bar{\epsilon})]$, and noting that $w(1) > 0, w(0) \leq 0$ and $w(\cdot)$ is continuous. The intermediate value theorem deliver λ^* where $w(\lambda^*) = 0$ and thus the interval for the second part is $(\underline{x}, \tilde{x} + \lambda\bar{\epsilon})$. □

We can now state the proof for the first part of Proposition 1 under scalar quality.

Proof. By contradiction, suppose that there exist two distinct elements in the identified set $(\gamma_0, f_0, \bar{\eta}_0)$, $(\gamma_1, f_1, \bar{\eta}_1) \in \mathcal{I}$. By the above lemma and assumption 2 in the main text, there exists two monotone partitional designs $\tilde{\psi}$ and ψ drawn with positive probability, such that: (1) for $\tilde{\psi}$ there is a partition $\tilde{r}$ where $\mathbb{E}_{f_0}[q|\tilde{r}, \tilde{\psi}] = \mathbb{E}_{f_1}[q|\tilde{r}, \tilde{\psi}]$; (2) for ψ there are two partitions r, r' such that $\mathbb{E}_{f_0}[q|r, \psi] < \mathbb{E}_{f_1}[q|r, \psi]$ and $\mathbb{E}_{f_0}[q|r', \psi] \geq \mathbb{E}_{f_1}[q|r', \psi]$. Where the directions of the inequality are assumed without loss. Using this, we have that

$$\gamma_0(\mathbb{E}_{f_0}[q|r, \psi] - \mathbb{E}_{f_0}[q|\tilde{r}, \tilde{\psi}]) = \eta_{rt} - \tilde{\eta}_{\tilde{r}} = \gamma_1(\mathbb{E}_{f_1}[q|r, \psi] - \mathbb{E}_{f_1}[q|\tilde{r}, \tilde{\psi}]) \implies \gamma_0 > \gamma_1$$
$$\gamma_0(\mathbb{E}_{f_0}[q|r', \psi] - \mathbb{E}_{f_0}[q|\tilde{r}, \tilde{\psi}]) = \eta_{r't} - \tilde{\eta}_{\tilde{r}} = \gamma_1(\mathbb{E}_{f_1}[q|r', \psi] - \mathbb{E}_{f_1}[q|\tilde{r}, \tilde{\psi}]) \implies \gamma_0 \leq \gamma_1$$

Contradicting that $(\gamma_0, f_0, \bar{\eta}_0)$ and $(\gamma_1, f_1, \bar{\eta}_1)$ are in the identified set. $\mathcal{I}$ is a singleton. □

The result under ignorance is straightforward as quality is bounded within the unit hypercube, and differences in WTP are identified. For the robust design, only the lower bound is relevant, which can obtain by taking the smallest difference in WTP across stars, $\underline{\Delta}\eta$, and defining $\underline{\Gamma} = \{\boldsymbol{\gamma} \in \mathbb{R}_+ | \boldsymbol{\gamma}'\mathbf{1} = \underline{\Delta}\eta\}$.

The extension to multiple dimensions of quality follows from the following lemma.

Lemma 2. *Let f, g be two distinct, continuous, strictly positive densities, supported on $[0,1]^n$. $\forall w \in \mathbb{R}^n_+$ strictly positive, $\exists a < b < c \in \mathbb{R}_+$ such that either $\mathbb{E}_f[w'x|w'x \in (a,b)] \geq \mathbb{E}_g[w'x|w'x \in (a,b)]$ and $\mathbb{E}_f[w'x|w'x \in (b,c)] \leq \mathbb{E}_g[w'x|w'x \in (b,c)]$ with one of the inequalities strict, or the analogous statement hold with the roles of f, g reversed. Also, there exists another $a, b \in \mathbb{R}_+$ such that $\mathbb{E}_f[w'x|w'x \in (a,b)] = \mathbb{E}_g[w'x|w'x \in (a,b)]$*

Proof. We can normalize one of $\boldsymbol{w}$ elements to 1 as all comparisons can be rescaled relative to it. Let $w_1 = 1$, and note that f and g define distributions over the random variable $y = \boldsymbol{w}'\boldsymbol{x}$.

To find the implied distribution, define the linear map $\boldsymbol{y} = (\boldsymbol{w}'\boldsymbol{x}, x_2, \ldots, x_n) = W\boldsymbol{x}$, where W is an identity matrix with its first row replace by $\boldsymbol{w}$. Using this and the standard change-of-variables method, the distribution of $\boldsymbol{y}$ induced by the prior f is

$$f_{\boldsymbol{y}}(\boldsymbol{y}) = \frac{1}{|det(W)|} f(W^{-1}\boldsymbol{y}) \quad \boldsymbol{y} \in \{\boldsymbol{y} \in \mathbb{R}_+ \times [0,1]^{n-1} : \sum_{i=2}^{n} w_i y_i \leq y_1 \leq \sum_{i=2}^{n} w_i y_i + 1\}$$

and so integrating over the marginals of $\boldsymbol{y}$

$$f_y(y) = \int_{V(y)} f(y - \sum_{i=2}^{n} w_i v_i, \boldsymbol{v}) d\boldsymbol{v} \quad V(y) = \{\boldsymbol{v} \in [0,1]^{n-1}, \boldsymbol{w}'\boldsymbol{v} \leq y\}$$

Where I used that $W = I + \boldsymbol{u}\boldsymbol{v}'$ where $\boldsymbol{u} = [1, 0, \ldots, 0]$ and $\boldsymbol{v} = [0, w_2, \ldots, w_n]$, which by the matrix determinant lemma implies that $|det(W)| = 1$. The implied distribution by g is defined analogously. Thus, we only need to prove that $f_y(\cdot)$ and $g_y(\cdot)$ satisfy the conditions of Lemma 1. First, note that both are positive and supported on $[0, \sum_{i=1}^{n} w_i]$. However, any closed interval of $\mathbb{R}$ is isomorphic to $[0,1]$, so continuity is the only property left to verify.

Let $\{y_n\} \in [0, \sum_i w_i]$ be a convergent sequence to y. Define the extension $\tilde{f} : [-\sum_{i=2}^{n} w_i, 1] \times [0,1]^{n-1} \to \mathbb{R}_+$ such that $\tilde{f}(\boldsymbol{x}) = f(\boldsymbol{x})$ if $x_i \geq 0$, and $\tilde{f}(\boldsymbol{x}) = 0$ otherwise. Note that $\mathbb{1}\{\boldsymbol{v} \in V(y)\} f(y - \sum_{i=2}^{n} w_i v_i, \boldsymbol{v}) \leq \tilde{f}(y - \sum_{i=2}^{n} w_i v_i, \boldsymbol{v})$ for every $y, \boldsymbol{v}$. Also, $\tilde{f}$ is integrable as f is. Continuity follows from the dominated convergence theorem. □

Finally, we can state the proof of proposition 1 in the multidimensional case.

Proof. Fix a scoring rule slope w. Lemma 2 delivers partitions used for baseline and contradiction. With this, obtain ψ and $\tilde{\psi}$ such that $\mathbb{E}_{f_0}[w'q|r,\psi] < \mathbb{E}_{f_1}[w'q|r,\psi]$ and $\mathbb{E}_{f_0}[w'q|r',\psi] \geq \mathbb{E}_{f_1}[w'q|r',\psi]$ and as before

$$\gamma_0'(\mathbb{E}_{f_0}[q|r,\psi] - \mathbb{E}_{f_0}[q|\tilde{r},\tilde{\psi}]) = \eta_{rt} - \tilde{\eta}_{\tilde{r}} = \gamma_1'(\mathbb{E}_{f_1}[q|r,\psi] - \mathbb{E}_{f_1}[q|\tilde{r},\tilde{\psi}])) \implies w'\gamma_0 > w'\gamma_1$$
$$\gamma_0'(\mathbb{E}_{f_0}[q|r',\psi] - \mathbb{E}_{f_0}[q|\tilde{r},\tilde{\psi}]) = \eta_{r't} - \tilde{\eta}_{\tilde{r}} = \gamma_1'(\mathbb{E}_{f_1}[q|r',\psi] - \mathbb{E}_{f_1}[q|\tilde{r},\tilde{\psi}])) \implies w'\gamma_0 \leq w'\gamma_1$$

Which delivers the contradiction. □

Connection to Abaluck and Gruber (2011): The demand estimates suggest consumers value a dollar in expected benefits similar to a \$2.5 reduction in premiums. As benefits are related to expected spending given each plan's cost-sharing attributes, this result appears opposite to the findings of Abaluck and Gruber (2011) for Medicare part D. This conflict, however, is artificial. First, the cost-sharing benefits computed by CMS are for a unit-risk enrollee in TM. They are therefore related not to the expected OOP spending Abaluck and Gruber use but to their "Cost Sharing" index. Their estimations (Table 1) find that consumers value this index more than premiums.[8] Second, CMS's computation of benefit ignores moral hazard in spending. As consumers are more generously subsidized in MA than in TM, the benefit metric is underestimated, which inflates the estimated coefficient. Finally, the level and the variance of spending in MA differs from that of Part D, making the comparison difficult. To evaluate my findings against estimates from a similar insurance market, I use the model and estimates from Handel *et al.* (2015) and CMS's OOP calculator to compute the value of every plan in the ten largest counties in Texas during 2015. I find that, on average, a \$1 increase in benefits for consumers equals a \$5.6 reduction in premiums, an even larger preference for benefits than in my findings.

Identification of investment risk: Quality in MA is the outcome of firm contracts with internal agents (e.g., preventive care staff) and external providers (e.g., hospital networks), interacting with local shocks to enrollee's compliance and health. I express this idea through a mapping between the quality and investment, $q_{ckt} = \Phi_k(x_{ckt} + \epsilon^M_{m(c)kt} + \epsilon^F_{f(c)kt})$. In this expression, $\Phi_k(\cdot)$ is a known strictly increasing function, $\epsilon^M_{m(c)kt}$ is the market-level shock, and $\epsilon^F_{f(c)kt}$ the firm-level shock.[9] Market distortions include a harsh flu season or a community vaccination drive affecting the performance of all insurers. Firm distortions capture events such as cost shocks to provider contract negotiations or firm-level congestion in following up with patients. Given this structure, identification follows from standard results in the non-parametric measurement error literature (Schennach, 2016).

Proposition 1. *(Identification of investment risk) Let $\boldsymbol{z}$ denote the vector of observables firms use to form beliefs about rival actions. Assume ϵ^M and ϵ^F are independent of each other, and $\boldsymbol{z}$, mean zero, symmetric, and have well-defined densities with nowhere-vanishing Fourier transforms. Then, the distributions of the errors are identified and given by*

$$f_{\epsilon^M_k}(\epsilon) = \mathcal{F}^{-1}\left(\mathbb{E}_z\left[\frac{|\mathcal{F}(f)_{\Delta^F \Phi_k^{-1}(q_k)|\boldsymbol{z}}(t)|^{1/2}}{|\mathcal{F}(f)_{\Delta^M \Delta^F \Phi_k^{-1}(q_k)|\boldsymbol{z}}(t)|^{1/4}}\right]\right); \quad f_{\epsilon^F_k}(\epsilon) = \mathcal{F}^{-1}\left(\mathbb{E}_z\left[\frac{|\mathcal{F}(f)_{\Delta^M \Phi_k^{-1}(q_k)|\boldsymbol{z}}(t)|^{1/2}}{|\mathcal{F}(f)_{\Delta^F \Delta^M \Phi_k^{-1}(q_k)|\boldsymbol{z}}(t)|^{1/4}}\right]\right) \quad \forall k$$

Where $\mathcal{F}(\cdot)$ is the Fourier transform, Δ^M is the within-market across-firms difference operator, and Δ^F is the within-firm across-market difference operator.

Proof. Denote $y_{jk} \equiv \Phi_k^{-1}(q_{jk})$ the normalized quality. The model implies that $y_{jk} = x_{jk} + \epsilon^M_{mk} + \epsilon^F_{fk}$, where x_{jk} is the unobserved investment. As this result is independent of the category, its index is omitted in what follows. Since the mean of x_j equals that of y_j, we

[8] To quote them, "Individuals are willing to pay a price in premiums for desirable plan characteristics, but this price is insufficiently sensitive to their individual circumstance" (Abaluck and Gruber, 2011).

[9] In estimation, I take $\Phi_k(x) = \Phi(x)(1 - \underline{q}_x) + \underline{q}_k$ where $\Phi(\cdot)$ is the standard normal CDF, and $\underline{q}_k$ is the minimum value of quality k that firms can produce. For example, an insurer cannot contract with a hospital to act in a way that would actively harm patients. The choice of $\Phi_k(\cdot)$ is not fundamental as it only provides an interpretation for the abstract unobserved investment.

can assume that all inputs are normalized to have mean zero. By the definition of z, $x_j|z$ is independent of $x_{j'}|z$ if different firms offer j and j'. Moreover, $y_j|z = x_j|z + \epsilon_j^M + \epsilon_j^F$.

Let (m, m') be two markets where firm f operates. Taking the within-firm difference we get, $\Delta^F y_f|z = \Delta^F x_f^*|z + \Delta^F \epsilon_{(m,m')}^M$. If there are multiple firms in (m, m'), this problem is analogous to a non-parametric unobserved measurement error problem with $N > 1$ repeated observations (Schennach, 2016). Applying the standard deconvolution to this class and using symmetry and mean-zero delivers $\mathcal{F}(f)_{\epsilon_{(m,m')}^M} = |\frac{\mathcal{F}(\Delta^F y_f|z)}{|\mathcal{F}(\Delta^M \Delta^F y_f|z)|^{1/2}}|^{1/2}$ where $\Delta^M \Delta^F y_f|z$ is the result of taking the difference of $\Delta^F y_f|z$ across two of the firms overlapping in (m, m'). Integrating the distribution of z on both sides and taking the inverse Fourier transform delivers the desired result. The result for the firm-level shock is analogous, taking first the difference within-market across firms. $\square$

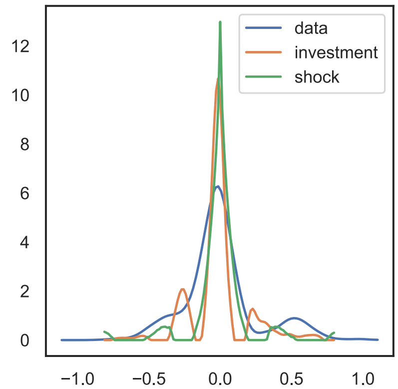
The intuition behind this result is that conditioning on firms' information set when investing is akin to conditioning on their investment choices. Consequently, any residual correlation in quality across firms within a market is driven by market-level quality shocks. Any residual correlation across markets within a firm is due to firm-level shocks.

Estimation of investment risk: The quality shock distributions are estimated non-parametrically by solving the conditional deconvolution outlined in the proof above.[10] In the set of observables used by firms, I include benchmarks, rebate fractions, bundled services, plan types, market sizes, and means, variances, and correlations between the same set of variables for rivals. Additionally, I include indicators for the presence of the top ten firms (by all-time enrollment) in the market. Overall, the vector contains over a hundred attributes observable by firms when investing. However, firms likely use only a few of these to form beliefs as rivals' targets are only relevant insofar as they affect demand.[11] This observation suggests a sparse relationship between quality and the conditioning set that the estimator leverages. The estimation proceeds in three steps.

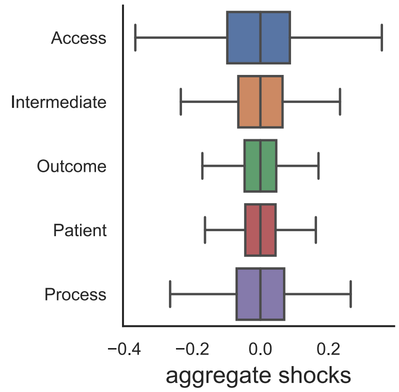
First, I assign *shock markets* to plans and form the data set used for estimation. I select markets with at least two firms that overlap in another market. Markets that fail to satisfy this condition are not helpful for estimation. Most contracts offered in these markets are found only in one market or have the vast majority of their demand in one. However, there is a fraction for which the assignment is less clear. To avoid specifying this manually and arbitrarily, I use the Resident-Matching algorithm with both markets' and contracts' matching preferences set according to observed enrollment. This way, these contracts are assigned to markets with less loss of valuable data. I then compute the observable market characteristics that enter the conditioning vector $\boldsymbol{z}$.

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

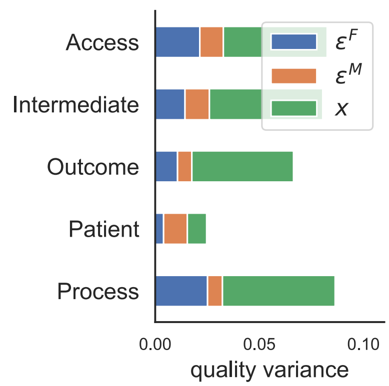
[11] For example, knowing that one competes against Humana, which systematically commands a significant market share, is probably enough to render the attributes of all smaller rivals irrelevant.



(a) Deconvolution



(b) Shock distributions



(c) Variance decomposition

Figure 6: Quality deconvolution

Note: Figure (a) displays the deconvolution of the Outcome category. Figure (b) shows the distribution of investment risk by category. Figure (c) shows the fraction of observed variance in quality attributed to investments and the two types of shocks.

Second, I used the data to estimate the conditional distribution of differences in realized quality (i.e., the distributions of $\Delta^F x_f|z$, $\Delta^M x_f|z$, $\Delta^M \Delta^F x_f|z$). To estimate the non-parametric conditional density, I use the estimator of Izbicki and Lee (2017). Specifically, I use the FlexZBoost implementation of Dalmasso *et al.* (2020), which combines this estimator with the XGBoost Machine Learning algorithm (Chen and Guestrin, 2016). I use a cosine basis with a maximum of 30 elements to minimize the mean squared error of predictions. I use a 20% sample of random permutations of the data to train the algorithm and the remainder to tune it. This estimator returns an estimate of each density at a collection of points. I ensure the symmetry of the distributions, as assumed in the theorem, by averaging the recovered density and its reverse-order values. Figure 6a illustrate how the deconvolution separates the observed quality distribution into investment and shock distributions.

This estimation recovers the ten distributions modeled: two shock types for five categories, illustrated in Figures 6b and 6c. The results show that shocks account for 39.6% of the variation in observed quality. Patients' assessments are the noisiest, as insurers cannot contract for better reviews, and, correspondingly, most of the variance is due to market-level shocks. Firm-level shocks are most important in Process measures, as they are insurer-labor intensive and involve monitoring patients and their care.

Firms' beliefs about rivals: Firms in the model hold rational expectations about rivals' unobserved investment actions. These investments, however, are not in the data. Nevertheless, each insurer is affected by its rivals' investments only to the extent that they affect realized quality, the distribution of which is observed. While we could identify and use the distribution of rival qualities to evaluate each firm's expectations, this would be challenging to compute. First, because rival qualities are not independent, and second because each firm would have to integrate over the realizations of hundreds of rival plans along five quality dimensions.

These challenges are alleviated using a lower-dimensional sufficient statistic for rivals' actions. Specifically, each firm is affected by its rivals' exclusively through their effect on the joint distribution of the individual preferences of each consumer i and the average of rival utility values $v_i = \frac{1}{|\mathcal{J}_{-f}|}\sum_{j\in\mathcal{J}_{-f}} \exp(\tilde{u}_{ij})$. Since the demand model has only a handful of consumer types, this distribution is of lower dimensionality and can capture rich correlation patterns across rival investments.

To implement this idea, I estimate the joint distribution of $(\alpha_i, \beta_i, \boldsymbol{l}_i, v_{if})$, where $\boldsymbol{l}_i$ is the consumers lock-in status vector. I consider only groups of (α_i, β_i) that are statistically different from one another in the demand estimates. Additionally, I group lock-in statuses into either "previously enrolled in MA", "previously enrolled in TM", or "new enrollee".[12] This results in 28 consumer types. Next, I use the demand estimates to compute v_{if} for each consumer, firm, and market and fit a lognormal probability distribution for each consumer type across markets.[13] I allow the parameters of the fitted distributions to depend on market characteristics. I use the same observables as when estimating investment risk, allowing the mean parameter to vary with the mean of these variables for rivals and analogously for the variance.[14] Finally, I combine the 28 marginals by fitting a Gaussian copula to the logarithm of rival values. I then draw 100 samples from the copula and use the inverse CDF of the fitted marginals to map these back to market-firm-level draws. At the end of the process, I obtain 100 independent draws for each firm-market and consumer type. These draws describe the estimated distribution, which I contrast with the aggregate rival-values data. The mean squared error of the fitted empirical CDF is $4.827 * 10^{-5}$.

It is worth noting, however, that the simplification of this approach comes at a cost. The dimensionality of v_i is smaller than that of the set of rival targets only if I do not condition on rivals' vector of non-quality attributes. However, the demand estimates indicate that quality, premiums, and benefits largely dominate v_i. The two last items are unknown to firms when investing in quality. Hence, the extent of firms' uncertainty goes beyond investments, which I capture through uncertainty over v_i. Nevertheless, the loss is in the connection between these beliefs and the equilibrium ensuing. Thus, this approach is an approximation that slightly extends the assumption regarding firms' beliefs to include uncertainty about other product characteristics determined after quality.

Investment costs: To estimate investment costs, I combine the estimated distribution of

[12] The insurer only uses v_{if} to predict the effect of changing quality on its future profits. This implies that dimensionality reductions of the parameter space are harmless as long as they do not systematically alter the likelihood that consumers will adopt or drop a plan as its quality changes.

[13] The data's distribution is remarkably similar to a log-normal with a mass-point near zero. The mass occurs when the firm is nearly a monopolist, facing only small rival plans. These markets are plentiful yet small in enrollment and their contribution to firms' profits.

[14] The average mean-square error (MSE) resulting from predicting the empirical cumulative distribution of rival values across all marginals was 0.00022.

investment risk, the empirical distribution of quality in the data, and firms' investment optimality condition. Next, I explain how I derive the regression equation used in the main text and how the conditional expectation of marginal revenue can be calculated from identified distributions. Throughout, I omit time indices and denote $\boldsymbol{y}_f = \Phi^{-1}(q_f)$, the mapping of observed quality back into investment space; y_{jk} is the expected investment of contract j in category k given the data. Additionally, I denote $\pi_f \equiv \sum_m \int \mathbb{E}[V_{fm}(\boldsymbol{q}_f, \boldsymbol{q}_{-f})]dF(\boldsymbol{q}_f|\boldsymbol{x}_f)$ the expected insurance profit of the firm.

Assuming, for now, the differentiability of π_f, we can consider the marginal investment profit of firms as a random variable in the data, and decompose it as $\frac{\partial \pi_f}{\partial x_{jk}} = \mathbb{E}[\frac{\partial \pi_f}{\partial x_{jk}}|\boldsymbol{y}_f] + \nu_{jkt}$. Where ν_{jkt} is the mean zero conditional on the observed $\boldsymbol{y}_f$. Rearranging and using firms' optimality conditions, we obtain the regression equation used in the main text.

Now note that we can re-express π_f as

$$\begin{aligned}
\pi_f &= \int \int \int V_f(\Phi(\boldsymbol{x}_f + \boldsymbol{\epsilon}_m + \boldsymbol{\epsilon}_f), \boldsymbol{q}_{-f}) f(\boldsymbol{q}_{-f}|\boldsymbol{\epsilon}_m, \boldsymbol{z}) f_{\epsilon_F}(\boldsymbol{\epsilon}_f) f_{\epsilon_M}(\boldsymbol{\epsilon}_m) d\boldsymbol{q}_{-f} d\boldsymbol{\epsilon}_m d\boldsymbol{\epsilon}_f \\
&= \int_{\epsilon_m} \int_{\epsilon_{-k}} \sum_{r \in R_{jk}(\boldsymbol{x}_j, \boldsymbol{\epsilon}_m, \boldsymbol{\epsilon}_{-k})} \int_{e_r(x_{jk})}^{e_{r+0.5}(x_{jk})} \tilde{V}_{fj}(\boldsymbol{x}_f + \boldsymbol{\epsilon}_m + \boldsymbol{\epsilon}_f|\boldsymbol{\epsilon}_m, r) f_{\epsilon_F}(\boldsymbol{\epsilon}_f) f_{\epsilon_M}(\boldsymbol{\epsilon}_m) d\boldsymbol{\epsilon}_m d\boldsymbol{\epsilon}_f
\end{aligned}$$

where $V_f = \sum_m V_{fm}$ is the sum of profits over markets, $\boldsymbol{z}$ are the observables that firm f uses to form beliefs about rival actions, $\boldsymbol{\epsilon}_m$ is the vector of market-level shocks in all markets and categories, and $\tilde{V}_{fj}(\cdot|\boldsymbol{\epsilon}_m, r) = \int V_f(\Phi(\cdot), \boldsymbol{q}_{-f}|r_j = r) f(\boldsymbol{q}_{-f}|\boldsymbol{\epsilon}_m, \boldsymbol{z}) d\boldsymbol{q}_{-f}$ is the firm's expected profit conditional on market shocks and a score, taking expectations over rivals' realizations. In the second equality, one dimension of firm-level shocks ϵ_{fk} was selected to partition the integral into segments that maintain the rating of x_{jk} constant. $R_{jk}(\boldsymbol{x}_j, \boldsymbol{\epsilon}_m, \boldsymbol{\epsilon}_{-k})$ is the set of ratings that plan j can reach through different firm-level shocks in category k, given the other values of integration. As the shocks have full support, this set includes all possible scores assigned by the regulator. The integration limits associated with each partition are denoted by $e_r(x_{jk})$ and equal to $-\infty$ and ∞ for $r = 1$ and $r = 5.5$, respectively.[15]

We can evaluate the derivative by using the Leibniz integral rule and the envelope theorem

$$\begin{aligned}
\frac{\partial \pi_f}{\partial x_{jk}} = &\int_{\epsilon_m} \int_{\epsilon_{f,-k}} \sum_{r \in R_{jk}(\boldsymbol{x}_j, \boldsymbol{\epsilon}_m, \boldsymbol{\epsilon}_{-k})} \Big(\tilde{V}_{fj}(\boldsymbol{x}_f + \boldsymbol{\epsilon}_m + [\boldsymbol{\epsilon}_{f,-k}, e_r(x_{jk})]|\boldsymbol{\epsilon}_m, r) f_{\epsilon_{Fk}}(e_r(x_{jk})) - \\
&\tilde{V}_{fj}(\boldsymbol{x}_f + \boldsymbol{\epsilon}_m + [\boldsymbol{\epsilon}_{f,-k}, e_{r+1}(x_{jk})]|\boldsymbol{\epsilon}_m, r) f_{\epsilon_{Fk}}(e_{r+1}(x_{jk})) \Big) f_{\epsilon_{F,-k}}(\boldsymbol{\epsilon}_{f,-k}) f_{\epsilon_M}(\boldsymbol{\epsilon}_m) d\boldsymbol{\epsilon}_m d\boldsymbol{\epsilon}_{f,-k} \\
&- \int_{\epsilon_m, \epsilon_f} \mathbb{E}[\sum_{m'} \gamma_{jm} D_{jm'}(\boldsymbol{p}_m(\boldsymbol{q}_m), r(\boldsymbol{q}_m))|\boldsymbol{q}_f = \Phi(\boldsymbol{x}_f + \boldsymbol{\epsilon}_m + \boldsymbol{\epsilon}_f), \boldsymbol{z}] \theta_k \phi(\boldsymbol{x}_f + \boldsymbol{\epsilon}_m + \boldsymbol{\epsilon}_f) dF(\boldsymbol{\epsilon}_m, \boldsymbol{\epsilon}_f)
\end{aligned}$$

[15]For interior values, the boundaries are equal to $e_r(x_{jk}) = \Phi_k^{-1}(\psi_k^{-1}(r - 0.25 - \sum_{k' \neq k} \psi_{k'}(\Phi_{k'}(x_{jk'} + \epsilon_{m(j)k'} + \epsilon_{f(j)k'})) - \omega_j) - x_{jk} - \epsilon_{f(j)k}$. The inverse ψ_k^{-1} might not be unique. In this case, for the lower integration limit, we take $\inf \psi_k^{-1}$ and the supremum for upper integration limits.

This expression, while complicated, has a simple interpretation. The first set of integrals corresponds to the change in profit due to increasing the probability of higher ratings and decreasing that of lower ratings. The second set of integrals is the change in profits associated with changing the marginal cost in the third stage. If service marginal costs were independent of quality, this term would be zero. Note that all of the densities involved in this equation are identified. The only unknown vector is $\boldsymbol{x}_f$.

The final step is in expressing $\mathbb{E}[\frac{\partial \pi_f}{\partial x_{jk}}|\boldsymbol{y}_f]$ as a function of identified densities. A change of variables shows that $\int_{-\infty}^{\infty} \frac{\partial \pi_f}{\partial x_{jk}}(\boldsymbol{x}_f) f(\boldsymbol{x}_f|\boldsymbol{y}_f) d\boldsymbol{x}_f = \int_{-\infty}^{\infty} \frac{\partial \pi_f}{\partial x_{jk}}(\boldsymbol{y}_f - \boldsymbol{e}) f_{\epsilon_F+\epsilon_M}(\boldsymbol{e}) d\boldsymbol{e}$ which can be evaluated given only the identified distribution of shocks and realized qualities.

Appendix III: Scoring Design

Empirical scoring design implementation: Each firm's insurance profit, $V_f(\psi, \boldsymbol{q})$ varies with $(\psi, \boldsymbol{q})$ only through their effect on scores, marginal costs, and the value that their rivals might offer in the market. Therefore, we can rewrite profits as $\tilde{V}_f(\boldsymbol{w}, \boldsymbol{c}, \boldsymbol{v})$ where $\boldsymbol{w} = \boldsymbol{\gamma}' \mathcal{E}[q|\psi(\boldsymbol{q})]$ is consumers' expected quality-utility, $\boldsymbol{c} = C(\boldsymbol{q}, \boldsymbol{z})$ are marginal costs, and $\boldsymbol{v}$ are rivals' value as defined in Appendix II (see Firms' beliefs about rivals). As the space of qualities is convex and compact and marginal costs are continuous functions of quality, $(\boldsymbol{w}, \boldsymbol{c}, \boldsymbol{v})$ lie on a convex compact set. Moreover, the implicit function theorem implies that $\tilde{V}_f$ is differentiable and, as marginal costs and demand are bounded, its first derivative is bounded. Hence $\tilde{V}_f$ is Lipschitz continuous, which implies strict bounds on the approximation error from linear interpolation within a finite grid on the space of $(\boldsymbol{w}, \boldsymbol{c}, \boldsymbol{v})$, and that as the number of grid points increases (uniformly over the domain), the linear interpolation of $\tilde{V}_f$ converges uniformly to the true value. This grid, $(\boldsymbol{w}, \boldsymbol{c}, \boldsymbol{v}) \mapsto \tilde{V}$, and its interpolation, is the key to simplifying the computation of the designer's problem. Crucially, the grid is computed only once, and each point can be solved independently in parallel.

Using the grid of equilibrium insurance profits, we can evaluate the total welfare of the market at any scoring rule ψ through the following steps. First, given ψ, compute consumers expected quality at each score and interpolate across the computed $\boldsymbol{w}$ to form a new grid $(\boldsymbol{w}(\psi), \boldsymbol{c}, \boldsymbol{v}) \mapsto \tilde{V}$. Second, initialize firms' beliefs about rivals to some initial distribution $G_f(\boldsymbol{v}_{-f})$ (e.g., the empirical distribution of baseline rival values). Third, find each firm's optimal investment choices. As $\tilde{V}_f$ is precomputed, this consists of solving

$$\max_{\boldsymbol{x}_f} \sum_{n=1}^{N_w \times N_c} \sum_{l=1}^{N_v} \tilde{V}_f(\boldsymbol{w}_n(\psi), \boldsymbol{c}_n, \boldsymbol{v}_l) \mathbb{P}_F(\boldsymbol{w}_n(\psi), \boldsymbol{c}_n | \boldsymbol{x}_f) G_f(\boldsymbol{v}_l) - I_f(\boldsymbol{x}_f, \boldsymbol{\mu}_f)$$

where $\mathbb{P}_F(\boldsymbol{w}_n(\psi), \boldsymbol{c}_n|\boldsymbol{x}_f)$ is the probability of expected quality-utility and marginal costs as determined by the investment choice and the estimated investment risk distribution, F. In the sum, N_d denotes the number of grid points over each dimension d in the grid. This problem is differentiable and can be solved independently for each firm in parallel. Third, update firms' $G_f(\boldsymbol{v}_{-f})$ to be consistent with these choices and iterate back to point two until convergence. Finally, I use the equilibrium investments to compute expected profits and the equilibrium bids (by-products of computing the grid of expected insurance profits) to compute expectations over consumers' surplus.

I make three simplifying assumptions when solving this procedure. First, I hold the lognormal shape restriction on firms' beliefs about rival actions fixed, changing only their means to satisfy rational expectations. Second, I hold consumers' prior fixed, only changing their posterior belief according to ψ. Third, I grid potential investment choices to accelerate finding optimal investments. I interpret the first two choices as an expression of the short-run effect of changing the scoring rules. Evaluating the long-run effects would involve specifying how a firm's belief distribution is formed and a process for adjusting consumers' priors. Additionally, it would require considering the dynamic effects of quality investment which is beyond this work's scope. The third choice is purely for computational convenience.

The methodological steps above make evaluating the designer's objective feasible. However, exploring the space of solutions is burdensome as the objective is not everywhere differentiable and features flat regions where changes in cutoffs or weights do not affect firms' choices. To find the global constrained optima of this function, I use the algorithm of Malherbe and Vayatis (2017). This algorithm is guaranteed to find the global optima of Lipschitz functions of unknown finite Lipschitz constant over a compact convex space with a non-empty interior.

Decomposition of monotone partitional scores: The empirical design methodology relies on a decomposition of monotone partitional scores into a polynomial aggregator and a cutoff function. Formally, a monotone partitional score is an injective mapping ψ from a convex and compact quality space $\mathcal{Q} \subset \mathbb{R}^n$ into a finite ordered set $(\mathcal{A}, \geq_A)$, such that $q \geq q' \implies \psi(q) \geq_A \psi(q')$. A polynomial aggregator is a polynomial function mapping $\mathcal{Q}$ to $\mathbb{R}$, and a cutoff function is a weakly monotonic step function from $\mathbb{R}$ to $\mathcal{A}$. The following proposition proves the decomposition.

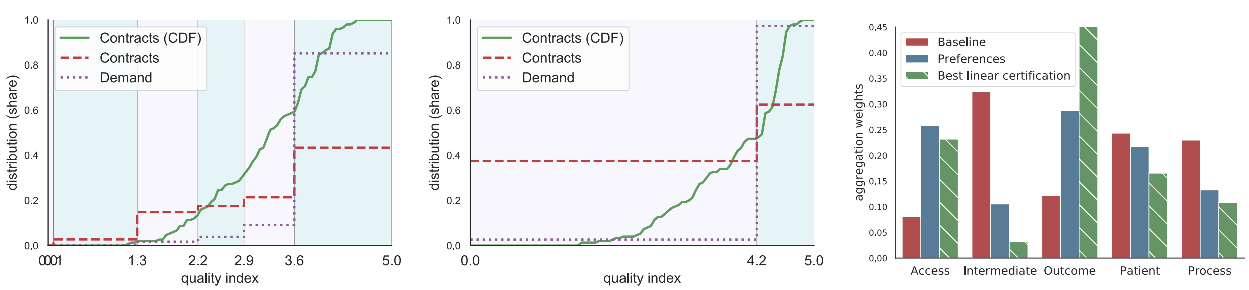
Proposition 2. *Let $\psi : \mathcal{Q} \to \mathcal{A}$ be a monotone partitional score. Then $\forall \epsilon > 0$, $\exists m > 0$, a polynomial aggregator of order m, P_m, and a cutoff function K_m such that $\psi^* = P_m \circ K_m$ satisfies $||\psi^* - \psi||_{L_1} < \epsilon$.*

Proof. Fix $\epsilon > 0$. Without loss of generality, $Q = [0,1]^n$ and $\mathcal{A} \subset [0,1]$. Assume $n > 1$ and that ψ takes on more than one value; otherwise, the proof is trivial. Every weakly monotone finite score partitions $[0,1]^n$ into a collection of finitely-many disjoint sets $\mathcal{M}$, such that ψ is constant over set $M \in \mathcal{M}$. Without loss, assume that the Lebesgue measure of every $M \in \mathcal{M}$ is positive; otherwise, the scoring set can be ignored under the L1 norm. Note that the boundary between every pair of contiguous sets is a set Δ_M of points, or equivalently, a line segment. Consider a small $\delta > 0$ such that the δ-neighborhood of each boundary $N(\Delta_M, \delta)$ do not overlap. Define the continuous function $f_\delta(x)$ as equal to ψ everywhere but in the boundary of the neighborhoods and equal to the linear interpolation between the steps of ψ across the boundary. Note that

$$||\psi - f_\delta||_{L1} = \int_{[0,1]^n} |\psi(x) - f_\delta(x)|dx = \sum_{\Delta_M} \int_{N(\Delta_M,\delta)} |\psi(x) - f(x)|dx < \sum_{x' \in M} [\psi(\Delta_M+) - \psi(\Delta_M-)]\frac{\delta}{2}$$

Where $\psi(\Delta_M+)$ and $\psi(\Delta_M-)$ are the values of ψ above and below the boundary. Now as f_δ is continuous, by the Stone-Weierstrass theorem there exists a polynomial P_m such that for all $x \in [0,1]$ we have $|P_m(x) - f_\delta(x)| < \frac{\epsilon}{2}$. Thus, picking $\delta < \frac{\epsilon}{\sum_{x' \in M}[\psi(\Delta_M+) - \psi(\Delta_M-)]}$ we have that $||P_m(x) - \psi|| = ||P_m - f_\delta + f_\delta - \psi|| \leq ||P_m - f_\delta|| + ||f_\delta - \psi|| < \epsilon$. Finally, note that we can pick K_m to shrink the approximation error further. $\square$

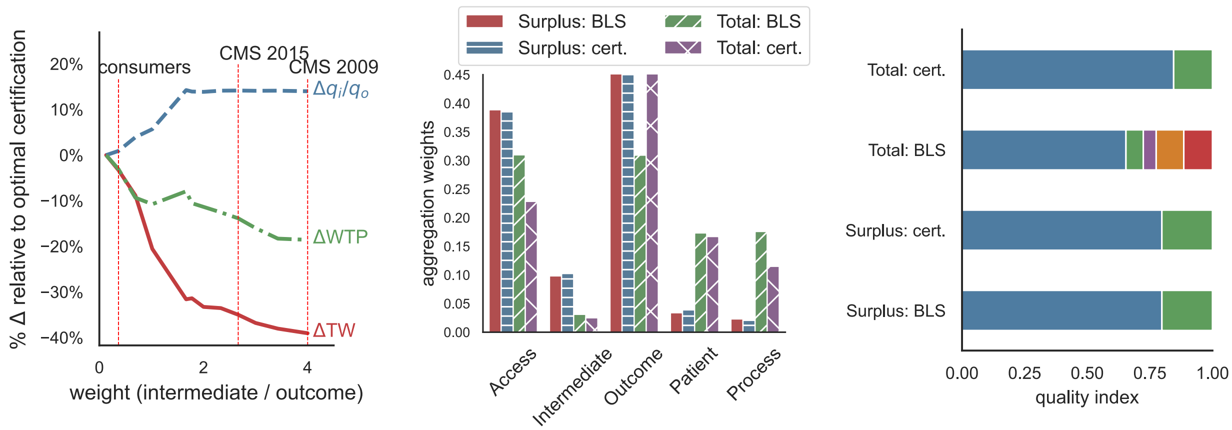
Alternative specifications: The main text presents the welfare value of two designs without showing their aggregator and cutoffs. The first is a design that holds CMS's weights fixed and optimizes cutoffs. These cutoffs, shown in Figure 7a, generate only five scores. Unlike the best linear substitute, cutoffs under this design are more evenly distributed on the quality index space, similar to CMS's original design. The second design is an optimal certification one. The optimal cutoffs and weights are shown in Figures 7b and 7c, respectively. This design is similar to the best linear substitute in construction and core mechanisms.



(a) CMS's optimal cutoffs (b) Cutoffs, certification (c) Weights, certification

Figure 7: Optimal design given CMS's weights and optimal certification

Regulatory preferences: The most significant difference between CMS's design and the best linear substitute is the relative weight allocated to the Intermediate Outcome category relative to the Outcome category. As discussed in the main text, various reasons might have led CMS to skew the design in this direction; hence, its value is ambiguous. The welfare cost, however, is measurable given the current estimates. To assess it, I optimize twelve certification designs, starting from the optimal weights, and shifting weight away from Outcome and into Intermediate, keeping their sum constant. The points are chosen to include the optimum, consumers' relative preferences for the categories, and the relative weight in every design CMS has tested in the data. Figure 8a shows the results. The blue dashed line shows that as the weights skew, the average Intermediate quality of contracts increases relative to Outcomes, as expected. However, the effect plateaus rapidly as consumers' WTP



(a) Regulatory preferences (b) Aggregation weights (c) Cutoffs

Figure 8: Regulatory preferences and design under alternative objectives

Note: Figure (a) displays the welfare change under optimal certification starting from the optimal weights and skewing the relative weights of the Intermediate Outcome and Outcomes categories to span consumers' preferences and all CMS's designs. Figures (b) and (c) show the optimal aggregation weights and cutoffs under the objectives of maximizing consumer surplus only or total welfare net of subsidies.

Table 7: Welfare effect of linear scores under different objectives

	Full Objective		**Consumer Surplus**	
	Certification	Substitute	Certification	Substitue
Δ Consumer Surplus	152.6	147.5	211.2	212.8
Δ Firm Profits	476.8	504.0	275.1	274.9
Δ Gov. Spending	-96.3	-81.5	-87.5	-87.4
Δ Total Welfare	629.4	651.4	486.2	487.8
MA share	56.5%	57.1%	48.1%	48.1%

Notes: This table displays the welfare effect of redesigning the scoring system, relative to the MA Star Rating baseline. The first two columns present the welfare effect of the best linear certification and substitute (nine scores) when the objective is to maximize total welfare net of government spending. The third and fourth columns maximize only consumer surplus. All values are in 2015 dollars per Medicare beneficiary.

for a certified product (green dashed-dotted line) drops rapidly, and thus firms' incentives to invest in quality deteriorate. Thus, the shift erodes information and quality; consequently, total welfare (solid red line) drops rapidly. Under the 2015 design, \$4.6 billion in welfare is lost among the markets used in the counterfactual. This, in turn, generates an added investment in the Intermediate category worth at most \$264 million, assuming uniform improvements across contracts.

Results for other objectives: Table 7 shows the welfare gains from optimal certification and the best linear substitute design when maximizing total welfare net of government spending (Full Objective) and when maximizing only consumer surplus. Figure 8 presents the designs, showing that they all feature pooling at the bottom, improved alignment with consumer preferences, a large shift in weights away from Intermediate Outcomes and towards Outcomes, and limited granularity. Interestingly, under a consumer surplus objective, certification is nearly optimal as it limits vertical differentiation across firms and controls prices at the expense of variety in quality.

References

Abaluck, J. 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.

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

Callaway, B. and Sant'Anna, P. H. (2020). Difference-in-Differences with multiple time periods. *Journal of Econometrics*, pp. 1–45.

Centers for Medicare and Medicaid Services (2016). Trends in Part C & D Star Rating Measure Cut Points. **2018**.

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, **13-17-Augu**, 785–794.

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.

Dalmasso, N., Pospisil, T., Lee, A. B., Izbicki, R., Freeman, P. E. and Malz, A. I. (2020). Conditional density estimation tools in python and R with applications to photometric redshifts and likelihood-free cosmological inference. *Astronomy and Computing*, **30** (Ml).

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

Handel, B., Hendel, I. and Whinston, M. D. (2015). Equilibria in health exchanges: Adverse selection versus reclassification risk. *Econometrica*, **83** (4), 1261–1313.

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

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

Malherbe, C. and Vayatis, N. (2017). Global optimization of Lipschitz functions. *34th International Conference on Machine Learning, ICML 2017*, **5** (1972), 3592–3601.

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

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