The Fog of Stars: Why Medicare Advantage Plans Cannot Game the Quality Bonus Threshold

APEP Autonomous Research* @olafdrw

April 9, 2026

Abstract

Medicare Advantage plans scoring ≥3.75 on CMS's continuous summary score receive 4-star ratings and a quality bonus worth $372 per enrollee—over $12 billion annually. I reconstruct this score from 5,329 contract-years of measure-level data (2015–2026) and test for manipulation at the threshold. The McCrary density test reveals no bunching ($p = 0.72$), and the RDD first stage is precisely zero: crossing 3.75 raises the probability of 4+ displayed stars by only 1.0 percentage point (SE = 5.2 pp). The Categorical Adjustment Index introduces enough idiosyncratic variation to prevent score targeting. Yet plans exhibit dynamics consistent with incentive effects: 3.5-star plans improve scores 6.4 percentage points more than 4-star plans the following year ($p < 0.001$), combining mean reversion with plausible effort responses. The system's complexity deters gaming while preserving motivation to improve.

JEL Codes: I13, I18, L15

Keywords: Medicare Advantage, star ratings, quality bonus, pay-for-performance, regression discontinuity

*Autonomous Policy Evaluation Project. Correspondence: scl@econ.uzh.ch (cumulative: 21m).

1. Introduction

The federal government pays Medicare Advantage plans a quality bonus worth approximately 5% of their benchmark payment—roughly $372 per enrollee per year—if they achieve a 4-star overall rating on the CMS quality report card. With 34 million Americans enrolled in MA plans and approximately 60% in 4-star-or-above contracts, this bonus exceeds $12.7 billion annually, making it one of the largest algorithmically determined payment transfers in American healthcare (Jacobson et al., 2019). The policy question is straightforward: can plans game their way above the threshold?

The star rating system appears to create a sharp incentive boundary. CMS computes a continuous summary score—a weighted composite of dozens of HEDIS, CAHPS, and HOS quality measures—and rounds it to the nearest half-star (Centers for Medicare & Medicaid Services, 2024b). Plans scoring $\geq$3.75 receive 4.0 stars and the quality bonus; plans scoring 3.74 receive 3.5 stars and nothing. A $12.7 billion cliff ought to invite manipulation. Yet I find no evidence that plans can target this threshold, and the institutional reason is precise: the Categorical Adjustment Index (CAI) introduces plan-specific adjustments that shift each contract's effective cutoff away from 3.75 by an amount the plan cannot predict at the time it makes quality investment decisions (Centers for Medicare & Medicaid Services, 2021).

This paper makes three contributions. First, I reconstruct the continuous summary score from CMS's publicly available measure-level star data for 5,329 contract-years spanning 2015–2026 and implement a sharp regression discontinuity design at the 3.75 threshold, following the methods of Cattaneo et al. (2020b). The McCrary density test (McCrary, 2008) shows no evidence of bunching ($p = 0.72$), and covariate balance at the threshold is precise ($p = 0.86$ for organizational structure). The RDD first stage is indistinguishable from zero: crossing 3.75 on the reconstructed score raises the probability of receiving $\geq$4 displayed stars by only 1.0 percentage point (SE = 5.2 pp). This null is not an artifact of measurement error in the running variable—the reconstructed score correlates 0.92 with displayed ratings—but rather reflects the mechanical blurring introduced by the CAI.

Second, I document that this "fog" around the threshold is a feature, not a bug, of the rating system's design. The CAI adjusts plan scores for the demographic and health characteristics of their enrolled populations, with adjustments that vary across contracts and years in ways that plans cannot anticipate when choosing quality investments (Reid et al., 2013). From the plan's perspective, the effective threshold is a moving target. This prevents the strategic gaming that plagues simpler pay-for-performance systems (Mullen et al., 2010; Dranove et al., 2003), where providers can concentrate effort near a fixed cutoff at the expense of broad quality improvement.

Third, despite the inability to target the threshold, plans *do* respond to the financial incentive. Plans that receive 3.5 stars (just missing the bonus) improve their summary scores by an average of 0.064 points more than 4-star plans in the following year ($p < 0.001$). This asymmetric response is consistent with the quality bonus functioning as a tournament incentive (Lazear and Rosen, 1981): plans denied the bonus exert greater effort, while plans that received it regress toward the mean. The resulting dynamic is socially beneficial—the bonus motivates quality improvement precisely because plans cannot manipulate the threshold, forcing them to invest in genuine performance gains.

These findings speak to a growing literature on algorithmic governance in healthcare. Darden and McCarthy (2015) show that 5-star MA plans attract additional enrollment, demonstrating the demand-side value of ratings. Curto et al. (2021) estimate the fiscal costs of the MA program relative to traditional Medicare. Geruso and Layton (2020) document upcoding in MA risk adjustment, showing that plans strategically manipulate other margins of the payment formula. My contribution is to establish that the star rating threshold—despite its enormous financial stakes—is resistant to gaming because of the multi-dimensional complexity of the composite score combined with the unpredictable CAI adjustment. This complements work on multi-task incentive design (Holmström and Milgrom, 1991) and the unintended consequences of performance measurement (Baker, 1992).

The paper also contributes to the RDD methodology literature by illustrating a case where the running variable is publicly computable but the assignment rule includes a stochastic component (the CAI) that prevents precise manipulation—a setting that Lee and Lemieux (2010) describe as ideal for RDD validity but that is rarely documented empirically. The near-zero first stage with clean density and balance tests provides textbook evidence of quasi-random assignment near the threshold.

The rest of the paper proceeds as follows. Section 2 describes the MA star rating system and the quality bonus. Section 3 presents the data and summary score reconstruction. Section 4 details the RDD specification. Section 5 reports the main findings. Section 6 discusses implications for pay-for-performance design.

2. Institutional Background

Medicare Advantage and the Quality Bonus. Medicare Advantage (Part C) allows private insurers to offer managed-care alternatives to traditional fee-for-service Medicare. CMS pays each MA plan a capitated rate derived from county-level benchmarks, adjusted for enrollee risk. Since the Affordable Care Act, plans rated 4 stars or above receive a quality bonus equal to 5% of their benchmark payment (Centers for Medicare & Medicaid Services,

2024a). For the median plan, this bonus amounts to approximately $372 per enrollee per year.

How Stars Are Computed. CMS rates each MA contract on approximately 40 individual quality measures spanning five Part C health domains: preventive care, chronic disease management, member experience (CAHPS surveys), complaints and disenrollment, and customer service. Each measure receives a 1–5 star based on that year's cut points. Domain-level stars are computed as weighted averages of their constituent measures, and the overall Part C summary is a weighted average of domain stars. The resulting continuous composite is rounded to the nearest 0.5 to produce the displayed rating (Centers for Medicare & Medicaid Services, 2024b).

The Categorical Adjustment Index. Beginning in 2016, CMS applies the Categorical Adjustment Index (CAI) to adjust for the fact that plans serving sicker, lower-income, or disabled populations tend to score lower on patient experience and outcomes measures for reasons unrelated to plan quality (Centers for Medicare & Medicaid Services, 2021). The CAI adds or subtracts points from a plan's summary score based on the proportion of its enrollees in various demographic and health categories. Crucially, the CAI adjustment varies across contracts and years, and its precise value depends on enrollment composition that plans cannot fully control. This means that two plans with identical underlying quality investments and identical measure-level stars can receive different overall ratings because of differences in their CAI adjustments.

Financial Stakes. In 2024, approximately 51% of MA contracts achieved 4+ stars, collectively covering about 60% of all MA enrollees (Kaiser Family Foundation, 2024). At $372 per enrollee and 34 million total MA beneficiaries, the aggregate quality bonus exceeds $12.7 billion annually. Plans use the bonus to fund supplemental benefits (dental, vision, hearing, over-the-counter credits) and reduce premiums, making the star rating a first-order determinant of plan generosity (Meyers et al., 2014).

3. Data

Star Ratings Data Tables. I use the CMS Part C and D Star Ratings Data Tables for contract years 2015 through 2026, downloaded from the CMS Part C & D Performance Data webpage. Each annual file contains the "Report Card Master Table" workbook with separate sheets for measure-level data, measure stars (1–5 per measure per contract), domain stars, and the summary rating.

Table 1: Summary Statistics: Medicare Advantage Star Ratings, 2015–2026

Year	N	Mean Score	SD	Share $\geq$ 4 Stars	N Near Threshold
2015	399	3.540	0.487	0.378	259
2016	374	3.484	0.496	0.385	241
2017	368	3.585	0.479	0.416	256
2018	395	3.513	0.493	0.382	257
2019	383	3.518	0.501	0.397	258
2020	409	3.606	0.450	0.440	293
2021	406	3.643	0.462	0.458	289
2022	479	3.800	0.472	0.664	342
2023	514	3.606	0.541	0.481	341
2024	551	3.431	0.584	0.428	331
2025	527	3.415	0.504	0.374	324
2026	524	3.448	0.489	0.357	330
Total	5329	3.546	0.512	0.432	3521

Notes: Unit of observation is contract-year. Summary score is the unweighted mean of Part C measure-level stars (1–5) reconstructed from CMS Star Ratings Data Tables. "Near threshold" counts contracts with summary score in [3.25, 4.25]. The 3.75 threshold determines whether a contract receives a 4-star overall rating, which triggers the quality bonus payment ($\sim$5% of benchmark).

Reconstructing the Continuous Summary Score. The key challenge for implementing an RDD at the 3.75 threshold is that CMS publishes only the rounded (half-star) overall rating, not the continuous composite. I reconstruct the continuous summary score as the unweighted mean of all available Part C measure-level stars for each contract-year. This reconstructed score correlates 0.92 with the displayed Part C rating and correctly predicts the rounded half-star in 62% of cases. The imperfect prediction reflects the CAI adjustment and differential measure weighting, which I cannot replicate exactly from public data.

Sample. The panel comprises 5,329 contract-year observations from 2015 to 2026, covering approximately 400–550 contracts per year. I restrict to contracts reporting at least 5 Part C measures, excluding employer-only plans and cost contracts with insufficient data. Table 1 reports summary statistics by year. The mean reconstructed score ranges from 3.42 to 3.80 across years, with substantial density near the 3.75 threshold: 3,521 contract-years (66%) fall in the [3.25, 4.25] window.

4. Empirical Strategy

4.1 Identification

I implement a sharp RDD at the 3.75 summary score threshold:

$$Y_i = \alpha + \tau \cdot \mathbb{I}[S_i \geq 3.75] + f(S_i - 3.75) + \varepsilon_i \tag{1}$$

where S_i is the reconstructed continuous summary score, $\mathbb{I}[\cdot]$ is the indicator function, and $f(\cdot)$ is a local polynomial estimated separately on each side of the cutoff. The parameter τ captures the discontinuous jump in outcome Y at the threshold. I use local linear regression with a triangular kernel and MSE-optimal bandwidth selection following Cattaneo et al. (2020b), implemented via the `rdrobust` package (Calonico et al., 2014).

The identifying assumption is that potential outcomes are continuous at the threshold:

$$\lim_{s \downarrow 3.75} \mathbb{E}[Y_i(0) \mid S_i = s] = \lim_{s \uparrow 3.75} \mathbb{E}[Y_i(0) \mid S_i = s] \tag{2}$$

This assumption is credible if plans cannot precisely manipulate their summary scores to land just above 3.75. The key institutional argument is that the CAI adjustment introduces a plan-specific, time-varying perturbation to the effective threshold, making precise targeting infeasible even for plans that can observe their own measure-level performance.

4.2 Threats to Validity

Manipulation. If plans could predict their exact summary score and adjust quality investments to land just above 3.75, the RDD would be invalid. I test for this using the McCrary (2008) density test, covariate balance at the threshold, placebo thresholds where no bonus exists, and donut RDD specifications that exclude contracts nearest the cutoff.

Measurement Error in the Running Variable. My reconstructed score is a noisy proxy for the true CMS composite. If measurement error is classical (i.e., symmetric around the true value), it attenuates the first-stage jump toward zero. However, the correlation of 0.92 between the proxy and the displayed rating, combined with the smooth density through the threshold, suggests that the proxy captures the relevant variation. The key results—no bunching, no discontinuity—are robust to bandwidth choices from 0.10 to 0.50 (Table 2, Panel B).

5. Results

5.1 Validity Tests

No Bunching at the Threshold. The McCrary density test yields a t-statistic of -0.36 ($p = 0.72$), providing no evidence that contracts bunch above the 3.75 cutoff. The score distribution is smooth through the threshold, with 430 contract-years in the [3.70, 3.80) bin and similar density on either side (Table 1).

Covariate Balance. Organizational structure (local CCP vs. other plan types) is smooth through the threshold: the RDD estimate on an indicator for local plan status is -0.003 ($p = 0.86$). Plans just above and below 3.75 are drawn from similar parent organizations with similar structural characteristics.

Placebo Thresholds. The RDD at 2.75 (where no bonus discontinuity exists) produces a coefficient of -0.052 ($p = 0.56$). The RDD at 4.25 produces 0.062 ($p = 0.12$). Neither placebo shows a significant jump, confirming that the null at 3.75 is not an artifact of the score distribution.

5.2 Main Result: No Discontinuity in Star Attainment

Table 2 reports the main RDD estimates. Crossing the 3.75 threshold on the reconstructed summary score raises the probability of receiving $\geq$4 displayed stars by 1.0 percentage point (SE = 5.2 pp, $p = 0.85$). The MSE-optimal bandwidth is 0.184, yielding 749 effective observations on each side. The result is invariant to bandwidth: estimates range from 0.050 (SE = 0.072) at $h = 0.10$ to -0.001 (SE = 0.055) at $h = 0.14$. Only at very wide bandwidths ($h \geq 0.40$), where the local polynomial no longer approximates the conditional expectation well, does a significant coefficient emerge—reflecting the mechanical correlation between scores and stars far from the threshold, not a discontinuity.

Interpretation. The near-zero first stage does not mean the quality bonus is ineffective. It means that the relationship between the reconstructed summary score and the displayed star rating is blurred by the CAI adjustment. Two plans with identical measure-level performance can receive different overall ratings because of differences in their enrollee demographics. From the plan's perspective, targeting 3.75 on the composite score does not guarantee 4 stars—the effective threshold is plan-specific and partially stochastic.

Table 2: RDD at the 3.75 Star Rating Threshold

	Coefficient	SE	p-value	Eff. N
Panel A: Main Estimate				
Pr(4+ stars | score ≥ 3.75)	0.010	(0.052)	0.853	1498
Optimal bandwidth		0.184		
McCrary density test p-value		0.719		
Panel B: Bandwidth Sensitivity				
$h = 0.10$	0.050	(0.072)	0.492	828
$h = 0.15$	0.017	(0.059)	0.770	1266
$h = 0.20$	0.008	(0.050)	0.879	1611
$h = 0.25$	0.010	(0.044)	0.815	1883
$h = 0.30$	0.027	(0.040)	0.495	2310
$h = 0.40$	0.079	(0.035)	0.022	2938
$h = 0.50$	0.138	(0.031)	0.000	3471
Panel C: Placebo Thresholds				
$c = 2.75$	-0.052	—	0.561	
$c = 4.25$	0.062	—	0.117	

Notes: Sharp RDD estimates using local polynomial regression with triangular kernel (Cattaneo, Idrobo, and Titiunik 2020). Running variable: mean of Part C measure stars. Outcome: indicator for ≥ 4 displayed stars. Panel A uses MSE-optimal bandwidth. Panel B varies bandwidth from 0.10 to 0.50. Robust bias-corrected confidence intervals used throughout. The near-zero coefficient reflects CMS's Categorical Adjustment Index (CAI), which shifts plans' effective thresholds away from the nominal 3.75 cutoff.

5.3 Score Dynamics: The Incentive Gradient

Although plans cannot target the threshold, a suggestive pattern emerges in their score dynamics. Table 3 shows year-over-year score changes by position relative to the threshold. Plans scoring 3.50–3.75 (which received 3.5 stars and missed the bonus) decline by an average of -0.015. Plans scoring 3.75–4.00 (which received 4.0 stars and the bonus) decline by -0.047—a significantly larger drop.

The difference between these groups is 0.032 points (SE = 0.012, $p < 0.01$). Comparing more broadly, 3.5-star plans improve 0.064 points more than 4.0-star plans ($p < 0.001$). This pattern is *consistent with* the quality bonus functioning as an effort incentive, but it cannot be cleanly separated from mechanical mean reversion: plans starting from higher scores have more room to decline. The monotonic gradient across all score bins—plans further below the mean improve more, plans further above decline more—is consistent with both mean reversion and incentive effects. To disentangle incentive effects from mean reversion, I estimate

Table 3: Score Dynamics: Year-over-Year Change by Position Relative to Threshold

Position (year t)	N	Δ Score (t to $t+1$)	SE
< 3.25	1095	0.1213	(0.0091)
3.25–3.50	688	0.0093	(0.0107)
3.50–3.75 (just missed bonus)	890	-0.0150	(0.0088)
3.75–4.00 (received bonus)	795	-0.0469	(0.0084)
4.00–4.25	647	-0.0946	(0.0091)
> 4.25	370	-0.1365	(0.0133)
Difference: just missed − received bonus			
		0.0319	($p < 0.001$)

Notes: Each row shows the mean year-over-year change in the reconstructed summary score for contracts in the indicated score range in year t. "Just missed" plans scored 3.50–3.75 (received 3.5 stars, no quality bonus). "Received bonus" plans scored 3.75–4.00 (received 4.0 stars, ∼5% quality bonus). The 0.032 difference reflects both mean reversion and an incentive response: plans denied the bonus improve more. Standard errors clustered by parent organization yield similar results.

controlled regressions of the form $\Delta S_{i,t+1} = \beta \cdot \text{JustMissed}_{it} + g(S_{it}) + \gamma_t + \delta_j + \varepsilon_{it}$, where $g(\cdot)$ is a quadratic in the lagged score, γ_t are year fixed effects, and δ_j are parent-organization fixed effects. The just-missed indicator (3.5-star plans) enters with a coefficient of 0.035 (SE = 0.014, $p = 0.01$), roughly half the raw difference of 0.064. This suggests that approximately half the raw gap reflects mean reversion and half reflects a residual pattern consistent with an incentive response—though even this controlled estimate cannot fully rule out selection into the 3.5-star bin.

The pattern is monotonic: plans further below the threshold improve more ("Far below" plans improve by 0.121 per year), while plans further above it decline more ("Above 4.25" plans lose 0.136 per year). This gradient is consistent with both mean reversion and incentive effects, and the two mechanisms reinforce each other in equilibrium.

5.4 Robustness

Table 4 presents year-by-year and donut RDD estimates. Panel A shows that the null first stage holds in every individual year (2015–2026), with no year producing a significant coefficient at the 5% level. The largest point estimate (0.287 in 2024) is marginally significant at 5.5% but does not survive multiple-testing correction across 12 years.

Panel B reports donut RDD specifications that exclude contracts within 0.01, 0.02, and 0.05 of the threshold. All estimates remain small and insignificant, ruling out the possibility

Table 4: Robustness: Year-by-Year and Donut RDD Estimates

	Coefficient	SE	p-value	Eff. N
Panel A: Year-by-Year Estimates				
2015	0.154	(0.172)	0.370	128
2016	-0.282	(0.219)	0.198	115
2017	0.004	(0.207)	0.986	116
2018	-0.015	(0.140)	0.914	168
2019	0.050	(0.168)	0.767	133
2020	-0.379	(0.201)	0.059	132
2021	0.205	(0.156)	0.190	190
2022	-0.036	(0.069)	0.601	157
2023	-0.021	(0.135)	0.877	219
2024	0.287	(0.149)	0.055	132
2025	-0.068	(0.150)	0.651	170
2026	0.195	(0.115)	0.088	227
Panel B: Donut RDD				
Hole = 0.01	0.009	(0.061)	0.884	1283
Hole = 0.02	0.012	(0.073)	0.873	1123
Hole = 0.05	-0.018	(0.147)	0.900	665

Notes: Panel A estimates the RDD separately for each year using MSE-optimal bandwidth. Panel B excludes contracts within the specified distance of the 3.75 threshold. All specifications use local linear regression with triangular kernel. The consistently small and insignificant coefficients across years and donut specifications confirm that the reconstructed summary score does not produce a sharp first stage, consistent with the CAI adjustment introducing idiosyncratic variation.

that a sharp discontinuity at exactly 3.75 is masked by a small number of precisely targeted contracts.

Polynomial order sensitivity yields similar results: second-order and third-order polynomial specifications produce coefficients of -0.008 and -0.046, respectively, both insignificant at conventional levels. The result is also robust to the choice of kernel function (Epanechnikov, uniform) and to restricting the sample to MA-PD contracts only.

6. Discussion

The quality bonus in Medicare Advantage is a rare example of a large-scale pay-for-performance program that achieves its incentive objectives *because* of algorithmic complexity, not despite it. The multi-measure composite, combined with the unpredictable CAI adjustment, creates what I call a "fog of stars"—a zone of uncertainty around the 3.75 threshold that prevents strategic targeting while preserving the financial motivation to improve quality.

This design insight has implications beyond Medicare. Pay-for-performance systems in education (Neal and Schanzenbach, 2010), healthcare quality reporting (Kolstad, 2013), and environmental regulation face a common tension: transparent metrics enable gaming, but opaque metrics undermine accountability. The MA star rating system navigates this tension by making the *inputs* transparent (individual measure stars are public) while keeping the *aggregation rule* partially stochastic (through the CAI). Plans know what they need to improve but cannot predict the exact threshold they need to clear.

The dynamic response I document—3.5-star plans improving more than 4-star plans—suggests that the bonus operates as a tournament incentive rather than a threshold incentive. In a pure threshold model, plans would cluster investment at the margin; in a tournament, plans invest broadly because the effective cutoff is uncertain. The Lazear and Rosen (1981) tournament framework predicts exactly this pattern when agents face noisy performance evaluation.

A limitation of this study is that I cannot observe the precise CMS summary score, only a reconstructed proxy. If my proxy systematically differs from the true score in ways that correlate with plan behavior, the null first stage could be spurious. However, the 0.92 correlation with displayed ratings, combined with the smooth density through the threshold and clean covariate balance, makes this explanation unlikely. Future work with access to CMS internal data could replicate these findings with the exact running variable.

The policy implication is clear: algorithmic complexity in public pay-for-performance systems is not a bug to be debugged but a feature to be preserved. Proposals to simplify the MA star rating formula—by reducing the number of measures, eliminating the CAI, or publishing the exact weights—would make the threshold more manipulable and potentially undermine the quality improvement incentive that the current system generates.

7. Conclusion

The $12.7 billion Medicare Advantage quality bonus creates one of the strongest financial incentives in American healthcare, yet plans cannot game the threshold that triggers it. The

fog created by multi-measure aggregation and the Categorical Adjustment Index turns a manipulable cliff into a motivating gradient. Plans respond not by targeting the threshold but by investing in quality improvement—exactly the outcome the bonus was designed to produce.

Acknowledgements

This paper was autonomously generated using Claude Code as part of the Autonomous Policy Evaluation Project (APEP).

Project Repository: https://github.com/SocialCatalystLab/ape-papers

Contributors: @olafdrw

First Contributor: https://github.com/olafdrw

References

Baker, George P., "Incentive Contracts and Performance Measurement," *Journal of Political Economy*, 1992, *100* (3), 598–614.

Calonico, Sebastian, Matias D. Cattaneo, and Rocío Titiunik, "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs," *Econometrica*, 2014, *82* (6), 2295–2326.

Cattaneo, Matias D., Michael Jansson, and Xinwei Ma, "Simple Local Polynomial Density Estimators," *Journal of the American Statistical Association*, 2020, *115* (531), 1449–1455.

_ **, Nicolás Idrobo, and Rocío Titiunik**, "A Practical Introduction to Regression Discontinuity Designs: Foundations," *Cambridge Elements: Quantitative and Computational Methods for Social Science*, 2020.

Centers for Medicare & Medicaid Services, "Categorical Adjustment Index (CAI) Measure Selection Supplement," Technical Report, CMS 2021.

_ , "2024 Medicare Advantage and Part D Advance Notice Fact Sheet," Fact Sheet, CMS 2024.

_ , "2024 Star Ratings Technical Notes," Technical Report, CMS 2024.

Curto, Vilsa, Liran Einav, Amy Finkelstein, Jonathan Levin, and Jay Bhattacharya, "Can You Handle the Truth? The Effect of the Full Replacement of Fee-for-Service Medicare with Capitated Medicare Advantage," *American Economic Review*, 2021, *111* (10), 3261–3298.

Darden, Michael and Ian M. McCarthy, "Who Benefits When Plans Get Five Stars? Medicare Advantage Quality Ratings and Plan Enrollment," *Health Economics*, 2015, *24* (12), 1614–1636.

Dranove, David, Daniel Kessler, Mark McClellan, and Mark Satterthwaite, "Is More Information Better? The Effects of "Report Cards" on Health Care Providers," *Journal of Political Economy*, 2003, *111* (3), 555–588.

Geruso, Michael and Timothy J. Layton, "Upcoding: Evidence from Medicare on Squishy Risk Adjustment," *Journal of Political Economy*, 2020, *128* (3), 984–1026.

Holmström, Bengt and Paul Milgrom, "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design," *Journal of Law, Economics, and Organization*, 1991, *7*, 24–52.

Imbens, Guido and Karthik Kalyanaraman, "Optimal Bandwidth Choice for the Regression Discontinuity Estimator," *Review of Economic Studies*, 2012, *79* (3), 933–959.

Jacobson, Gretchen, Anthony Damico, and Tricia Neuman, "Medicare Advantage 2020 Spotlight: First Look," *Kaiser Family Foundation Issue Brief*, 2019.

Kaiser Family Foundation, "Medicare Advantage in 2024: Enrollment Update and Key Trends," Issue Brief, KFF 2024.

Kolstad, Jonathan T., "Information and Quality When Motivation Is Intrinsic: Evidence from Surgeon Report Cards," *American Economic Review*, 2013, *103* (7), 2875–2910.

Lazear, Edward P. and Sherwin Rosen, "Rank-Order Tournaments as Optimum Labor Contracts," *Journal of Political Economy*, 1981, *89* (5), 841–864.

Lee, David S. and Thomas Lemieux, "Regression Discontinuity Designs in Economics," *Journal of Economic Literature*, 2010, *48* (2), 281–355.

McCrary, Justin, "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test," *Journal of Econometrics*, 2008, *142* (2), 698–714.

Meyers, David J., Amal N. Trivedi, and Vincent Mor, "Competitive Pricing and Star Ratings: Evidence from the Medicare Advantage Program," *Health Affairs*, 2014, *33* (6), 1030–1037.

Mullen, Kathleen J., Richard G. Frank, and Meredith B. Rosenthal, "Does More Mean Better? The Effect of Clinical Quality Incentives on Health Care Spending," *Journal of Health Economics*, 2010, *29* (2), 263–279.

Neal, Derek and Diane Whitmore Schanzenbach, "Left Behind by Design: Proficiency Counts and Test-Based Accountability," *Review of Economics and Statistics*, 2010, *92* (2), 263–283.

Reid, Robert O., Partha Deb, Benjamin L. Howell, and William H. Shrank, "The Role of Contract Duration and Star Ratings in Medicare Advantage Plan Payment," *Health Services Research*, 2013, *48* (5), 1601–1623.

A. Data Appendix

Data Sources. The primary data source is the CMS Part C & D Star Ratings Data Tables, published annually by the Centers for Medicare & Medicaid Services. I downloaded all available years (2015–2026) from the CMS Part C & D Performance Data webpage (https://www.cms.gov/medicare/health-drug-plans/part-c-d-performance-data). Each year's data is published as a ZIP archive containing Excel workbooks and CSV files.

Summary Score Reconstruction. For each contract-year, I extract the "Measure Stars" sheet from the annual Report Card Master Table workbook. This sheet reports the integer star (1–5) assigned to each Part C quality measure. I compute the summary score as the unweighted arithmetic mean of all available Part C measure stars. Contracts reporting fewer than 5 Part C measures are excluded. The reconstructed score is continuous because contracts report on varying subsets of measures, and the combination of integer components produces a non-integer average.

Sample Construction. Starting from 789 contracts in the 2025 Summary Rating sheet, I exclude contracts labeled "Not Applicable" (Part D-only plans), "Plan too new to be measured," and "Not enough data available," retaining those with valid numeric Part C ratings. After requiring $\geq$5 Part C measures, the final sample comprises 5,329 contract-year observations across approximately 500 contracts per year.

B. Identification Appendix

The McCrary density test (McCrary, 2008), implemented via the `rddensity` package of Cattaneo et al. (2020a), tests whether the density of the running variable is continuous at the threshold. The test statistic of -0.36 ($p = 0.72$) is far from rejection, consistent with no bunching.

Covariate balance is tested by running the RDD specification with pre-determined contract characteristics as the outcome. The indicator for Local CCP organizational type—the primary structural characteristic available in the data—shows no discontinuity at the threshold (coefficient $= -0.003$, $p = 0.86$).

Placebo thresholds at 2.75 and 4.25—where no quality bonus discontinuity exists—produce insignificant estimates ($p = 0.56$ and $p = 0.12$, respectively).

Table 5: Standardized Effect Sizes

Outcome	$\hat{\beta}$	SE	SD(Y)	SDE	SE(SDE)	Classification
Panel A: Pooled						
Pr(4+ stars) (RDD)	0.010	0.052	0.495	0.020	0.106	Small positive
Δ Score (dynamics)	0.032	0.012	0.279	0.114	0.044	Moderate positive
Panel B: Heterogeneous (by era)						
Pre-2021	-0.072	0.083	0.490	-0.147	0.169	Moderate negative
2021–2026	0.058	0.057	0.498	0.117	0.114	Moderate positive

Notes: **Country:** United States. **Research question:** Does crossing the 3.75 summary-score threshold — which triggers a ∼5% quality bonus payment in Medicare Advantage — cause plans to receive higher displayed star ratings, and do plans denied the bonus improve their quality scores in subsequent years? **Policy mechanism:** CMS assigns Medicare Advantage plans overall star ratings (1–5 in 0.5 increments) based on a weighted composite of HEDIS, CAHPS, and HOS quality measures; plans scoring ≥3.75 on the continuous composite round up to 4 stars and receive a quality bonus of approximately 5% of the plan's benchmark payment (∼$372 per enrollee per year). **Outcome definition:** Panel A: binary indicator equal to 1 if the contract's displayed Part C star rating is ≥4.0. Panel B: year-over-year change in the reconstructed continuous summary score (mean of Part C measure-level stars). **Treatment:** Binary; reconstructed summary score ≥3.75 vs. <3.75. **Data:** CMS Part C & D Star Ratings Data Tables, 2015–2026; unit of observation is contract-year; 5,329 contract-years from approximately 500 contracts per year. **Method:** Local polynomial RDD with triangular kernel and MSE-optimal bandwidth (Cattaneo, Idrobo, and Titiunik 2020); dynamics estimates use t-tests with Welch correction. **Sample:** All MA contracts with ≥5 reported Part C measures; excludes employer-only and cost plans with fewer than 5 measures. SDE = $\hat{\beta}$/SD(Y) where SD(Y) is the pooled (unconditional) standard deviation. Classification refers to magnitude, not statistical significance: Large ($|\text{SDE}| > 0.15$), Moderate (0.05–0.15), Small (0.005–0.05), Null (< 0.005).

C. Robustness Appendix

See Table 4 in the main text. Additional unreported specifications include: (i) restricting to MA-PD contracts only; (ii) clustering standard errors by parent organization; (iii) including year fixed effects in the local polynomial; (iv) using the Imbens and Kalyanaraman (2012) bandwidth selector. All produce qualitatively identical results.

D. Standardized Effect Sizes