Systematic Differences Among Third-Party-Graded Morgan Dollar Populations

Evidence from a Large Dealer-Sourced Image Corpus

Preprint | Version 1.2 | 22 August 2026

Troy Lowry

Independent Researcher

Abstract

Background: Third-party certification grades are central to the market for Morgan dollars, yet claims that one grading service is systematically stricter than another are usually based on reputation, anecdote, or small crossover samples. This study asks a narrower and directly observable question: whether the current populations of same-grade coins in different services' holders differ visually when measured by one common instrument.

Methods: Classic Morgan-dollar images were sourced directly from coin dealers and evaluated with the proprietary computer-vision grading system that underlies the CoinAiLyzer application. The model was used as a common measurement instrument, not as ground truth. Five-fold out-of-fold predictions were aggregated at the physical-coin level. After identity repair, exclusion of 148 physical coins with fold leakage, and the final preregistered classic-date scope correction, the primary cohort contained 121,119 image-level predictions representing 65,092 physical coins. The primary contrast was PCGS versus NGC within stated grade. Prespecified measures included raw score gaps, Cohen's d, probability of superiority, physical-coin bootstrap confidence intervals, direct grade-shape contrasts, date/mint controls, confidential source controls, same-coin old-versus-new replication, side-specific checks, probability-vector diagnostics, and a both-sides-present sensitivity analysis. After confirmatory closeout, separately frozen exploratory specifications examined within-service dispersion, issue-specific service-population outliers, absolute same-label dispersion and random-pair differences, and finally whether the center of a fixed stated grade shifts across date/mint issues within PCGS and within NGC. All exploratory issue analyses reused the same 207 year x mint x grade cells with at least 30 physical coins per service.

Results: The primary confirmatory cohort met the preregistered Strong Replication definition. At AU-55, PCGS coins scored 0.73 Sheldon points higher on average than NGC coins of the same stated grade (d=0.227; probability of superiority=56.4%). At AU-58, the gap was 0.79 points (d=0.304; probability of superiority=58.7%). Both effects exceeded AU-50 and the MS-63 through MS-66 commercial core on direct bootstrap contrasts, and MS-62 through MS-66 met the preregistered practical-parity criterion in the primary population. A smaller PCGS advantage reappeared at MS-67. A fully held-out classic external cohort of 27,272 physical coins reproduced upper-AU separation and MS-62 through MS-66 convergence, but did not reproduce the primary cohort's immediate decline after AU-58: its largest effect occurred at MS-60 (ensemble d=0.542). Thus, the narrow AU-58-peak description is cohort-dependent, whereas elevated separation around the upper-AU/low-Mint-State boundary followed by

convergence by MS-62 is stable across cohorts. In exploratory follow-ups, overall within-cell dispersion was nearly identical between services and no issue-specific service outlier survived BH FDR at $q<0.05$. The equal-cell median within-label SD was 0.63 Sheldon points for PCGS and 0.65 for NGC; in the median eligible cell, two random same-label coins differed by at least 1.0 modeled point with probability 0.24 and 0.26. A limited repeat-image diagnostic (63 same-coin, same-side units) had a median absolute re-score difference of 0.26 points, smaller than typical mid-Mint-State differences between distinct same-label coins but too sparse for formal error decomposition. Commercial Mint State date/mint averages were comparatively stable: at MS-65, 38 eligible issue means spanned 0.74 points in PCGS and 0.71 in NGC, with an SD of 0.16 in each service.

Conclusions: The current dealer-sourced holder populations do not differ by a single constant service offset. The primary cohort localized its largest separation at AU-55 and AU-58; the external cohort reproduced upper-AU separation but peaked at MS-60. Across both cohorts, the most stable description is elevated separation around the upper-AU/low-Mint-State boundary, convergence through much of MS-62 to MS-66, and a smaller upper-MS return at MS-67. Same-year, same-mint, same-service, same-grade coins are not visually interchangeable under the common ruler, providing quantitative support for “buy the coin, not the holder.” The limited repeat-image check indicates that ordinary re-imaging/model wobble is smaller than the observed same-label coin-to-coin spread in commercial Mint State, although the present repeat sample cannot determine exactly how much of that spread is coin quality versus measurement variability. Within the well-populated commercial Mint State issues studied here, average score levels are relatively stable across date/mint compared with the larger coin-to-coin spread within an issue. This argues against a simple claim that every Morgan issue operates on a substantially different visible Mint State standard. It does not establish an abstract issue-independent standard, because the model learned from professional labels and may inherit industry issue conventions. Weak strike was not independently measured; an issue-blind design with independent strike measurement is required to test whether generic standards are applied differently to characteristically weak-struck or scarce issues.

Core interpretive rule
A higher model score for one service at a fixed stated grade is evidence that the observed population in that service's holders looks stronger to the common measurement instrument. It is not proof that the other service graded incorrectly, that the model knows the objectively correct grade, or that the difference was caused by the original grading decision.

1. Introduction

Third-party grading transformed the market for United States coins by placing an expert-assigned grade and authentication judgment into a standardized holder. For Morgan dollars, a one-point change can materially alter market value, and the service name on the holder can affect liquidity and price. Collectors and dealers therefore debate whether leading services apply meaningfully different standards.

The conventional evidence is difficult to evaluate. Dealer experience is extensive but selected. Crackout and crossover anecdotes compare the same coin across holders but are usually few in number and often published because the outcome was notable. Population reports count assigned grades but do not describe the visual-quality distribution within a grade.

A large image corpus permits a different test. Coins carrying the same stated grade from different services can be evaluated with exactly the same model-based measurement system. The model need not be accepted as the final authority on grade. It must provide a stable common ruler that is applied identically, out of fold, and without explicit certification-service metadata. The scale-free rank statistic is particularly useful because global calibration error or monotonic compression cannot by itself create a within-grade service rank shift.

1.1 Related work and literature search

The Sheldon numerical scale originated as a condition-and-value framework for early American cents and was later adapted into the general 1-to-70 language used by United States grading services. Current PCGS and NGC materials publish generic numerical descriptions that incorporate wear, surface preservation, luster, and strike, while

acknowledging that grading remains an expert visual judgment [1,2,5]. The present study does not treat either service label as an independent physical reference standard.

Computer-vision research in numismatics has historically concentrated on coin attribution, classification, segmentation, legend reading, die matching, and reconstruction [6-8]. Recent work has begun to address numerical condition grading directly, including feature-engineered systems and dual-image or vision-language models for fine-grained Mint State classification [9-11]. Those studies primarily evaluate agreement with labeled grades or grade tolerance. They do not ask whether large current populations in different certification services occupy different visual distributions at the same stated grade.

A targeted search of scholarly web indexes and publisher sites on 22 August 2026 used combinations of “coin grading,” “automated coin grading,” “computer vision,” “Sheldon scale,” “PCGS,” “NGC,” and “Morgan dollar.” The search identified automated-grading and broader numismatic-vision literature but no published image-level study comparable in scale and design to the present same-label PCGS-NGC population comparison. This was a targeted literature search rather than a formal systematic review; the novelty claim is therefore limited to the search conducted, not an assertion that no related work exists.

Primary research question
Among physical classic Morgan dollars currently residing in holders of the same stated numerical grade, do visual-score distributions produced by one fixed out-of-fold computer-vision measurement system differ systematically by certification service after physical-coin aggregation and controls for date, mint, and confidential dealer/source composition? If so, does an improved version of the measurement system reproduce the grade-specific shape observed before its results were examined?

The discovery analysis using the prior grader suggested a specific, risky pattern rather than a generic PCGS advantage: modest separation at AU-50, materially larger separation at AU-55 and AU-58, a decline through MS-60 to MS-62, and much smaller effects through MS-63 to MS-66. MS-67 and MS-68 were treated as a separate upper-Mint-State question. This shape was frozen before the new-grader output was examined.

2. Study Design and Inferential Target

This is an observational comparative study of current certified-holder populations. Coins were not randomly assigned to grading services. The observed slab population has been shaped by submission choices, crossovers, crackouts, reconsiderations, resubmissions, and time. These facts limit causal claims about original grading-service behavior but do not eliminate the market relevance of the resulting populations.

The study tests	The study does not establish
Whether same-label service populations differ under one common measurement instrument.	Which service has assigned the objectively correct grade.
Where service differences are largest, smallest, or change direction.	That a service is wrong on an individual coin.
Whether differences survive date/mint controls, source controls, side-specific checks, identity restriction, a new grader, and a held-out cohort.	That any difference is caused solely by initial grading thresholds.
Whether the current holder populations differ in a way relevant to a buyer drawing from those populations.	That a service deliberately protects a grade or that any specific mechanism has been proven.

The confirmatory specification v2.0 was frozen at 08:44 EDT on 21 August 2026 and registered on OSF at 09:58:41 EDT, before the canonical new-grader OOF file was generated and before any new-grader PCGS-NGC result was inspected. The registration states that the old grader generated the hypothesis and that the new grader provides confirmatory measurement replication. This separation of prediction from later exploration follows the rationale for preregistration described by Nosek and colleagues [17,18].

3. Corpus, Identity, and Cohort Construction

3.1 Source and scope

All images in the analytic corpus were sourced directly from coin dealers. Dealer identities, source-specific acquisition details, and source labels are confidential. Internal source identifiers were retained solely to test and control dealer/photography-source effects.

The confirmatory scope was restricted before result review to classic Morgan dollars dated 1878-1904 or 1921. A final closeout audit discovered that 157 image rows representing 91 physical modern Morgan issues dated 2021-2025 had nevertheless remained in the clean analytic file. This was an implementation error, not a new exclusion decision: the frozen specification explicitly excluded modern issues. The scope rule was therefore applied mechanically after documenting and hashing the omission. No physical identity mixed classic and modern years. Only two modern physical coins entered AU-50 through MS-67 (one NGC MS-65 and one PCGS MS-66), and the corrected classic-only estimates changed by no more than 0.001 in raw gap, d, or PoS. The v9 training and OOF cohort otherwise remained the frozen 18 August cohort; a later pipeline repair that recovered additional archive records was not added because doing so would have required rebuilding folds and retraining after the specification was frozen.

3.2 Final primary cohort

Corpus characteristic	Final confirmatory value
Canonical OOF image rows before physical-leakage exclusion	121,833
Canonical OOF fold-group identifiers before identity repair	65,341
Clean primary image rows after excluding 148 leaked physical coins	121,276 before classic-scope correction
Clean primary physical-coin identities	65,183 before classic-scope correction
Cert-anchored identities in pre-scope clean cohort	57,979 (sensitivity cohort later re-filtered to classic scope)
coin_key-fallback identities in pre-scope clean cohort	7,204 (sensitivity bookkeeping before classic-scope correction)
External CV-gate sensitivity image rows	50,213 classic-only (90 modern rows removed)
External CV-gate sensitivity physical coins	27,272 classic-only (62 modern coins removed)
Modern-issue rows removed by registered scope correction	157 rows / 91 physical coins
Final classic-only primary image rows	121,119
Final classic-only primary physical coins	65,092

The primary OOF cohort is a CV-verifiable subset of the eligible dealer-sourced corpus. At the row level, the verification gate retained 79.2% of eligible NGC images and 74.6% of eligible PCGS images. The gap was 1.3 percentage points at AU-55, 4.6 points at AU-58, and 14.3 points at MS-67. This service-dependent selection was not ignored. It motivated a separately frozen external sensitivity cohort containing records with neither a CV agreement nor a CV mismatch. All physical overlap with the v9 train, validation, and test corpus was removed before external scoring.

3.3 Physical-coin identity and leakage audit

The v9 training primitive pid was a bare product_id and was not always a physical-coin identifier. A pre-score audit found 171 product-ID collisions and 183 repeated service-certification identities represented under multiple product IDs or coin keys. The final analysis_coin_id therefore used a service-consistent certification identity when available and coin_key otherwise. Thirty rows with disagreement between the manifest certifier and the service prefix embedded in the extracted certification field were assigned by coin_key rather than certification number.

A second audit checked whether the repaired physical identity crossed folds. It identified 148 physical coins represented under multiple fold-group IDs and therefore present in the training data for a model used to produce another representation's nominal OOF prediction. All 148 identities and 557 associated image rows were excluded before service results were inspected. The final clean cohort therefore satisfies physical-coin OOF rather than merely PID-level OOF.

The side-specific analysis was also audited. The original side table used image rows directly. Only 63 coins, approximately 0.06% of side observations, contributed two images of the same side. A frozen corrective analysis averaged repeated same-side images before computing means, d, and PoS. Point estimates changed only in the fourth decimal, but the registered physical-coin unit was restored.

Upstream corpus construction also used exact perceptual-hash (pHash) matches to remove apparent duplicate listing images. A prepublication manual audit of a frozen random sample of 100 Morgan records dropped by this step found 17 correct duplicate removals, 53 false merges, 1 ambiguous case, and 29 cases whose counterpart could no longer be resolved from the current manifest. The false merges arose primarily from template-dominated composite listings and caused legitimate records to be dropped rather than transferring another coin's metadata into a retained observation. Extrapolation within the affected archive source suggests roughly 4,800 Morgan rows may have been unnecessarily removed as coverage loss. The retained analytic corpus is therefore not contaminated by these false merges, but coverage of that archive source is reduced. The audited drops showed no meaningful certification-service direction bias.

3.4 External held-out cohort

The external sensitivity cohort comprised stage-3 eligible records whose product ID had no CV mismatch and no CV agreement. These records were excluded before v9 training. The original frozen external set contained 27,334 physical coins and 50,303 image rows after removal of 52 identities overlapping v9 train, validation, or test. The final classic-scope audit removed an additional 62 modern physical coins represented by 90 image rows, leaving 27,272 classic physical coins and 50,213 image rows with zero physical-coin overlap with the v9 corpus. Each external obverse was scored by all five frozen obverse fold models and each reverse by all five frozen reverse fold models, with no retraining. The external analysis is a sensitivity test for CV-gate selection and out-of-training-corpus generalization, not a second independently collected random sample.

4. Common Model-Based Grading Instrument

4.1 Role and commercial provenance

The proprietary computer-vision grading system used in this study is the grading engine underlying the CoinAiLyzer commercial application. It was developed for automated coin grading rather than for the present service-comparison study. Throughout this paper it is treated as a measurement instrument, not as an oracle or a definition of the correct grade.

The author has a financial interest in CoinAiLyzer and in the proprietary grading technology used as the study instrument. That relationship is disclosed explicitly because the instrument is part of a commercial product.

4.2 Confidentiality boundary

The paper discloses the OOF design, physical-coin aggregation, score scale, probability-vector function, relevant validation checks, the service-comparison statistics, and the methodological fact that the grading input is isolated from the surrounding holder/background before inference. It does not disclose model architecture, feature design, the implementation details of proprietary preprocessing, loss functions, training recipe, or other commercially useful construction details.

4.3 Five-fold out-of-fold evaluation

The new grader retained ten fold-specific models, five obverse and five reverse, together with exact split files, class mappings, training logs, and original validation predictions. Canonical fp32 probability vectors were regenerated by inference only from the frozen fold weights. Verification confirmed exactly-one OOF coverage, train/validation PID disjointness, single-split assignment of every PID across sides and repeat photographs, identical class mappings across all ten folds, probability sums near one, and high argmax agreement with the original bf16 validation predictions. The subsequent physical-identity audit removed the 148 cross-PID physical leaks described above.

4.4 Output scale and rank invariance

For each image, the model produces a 20-class probability distribution on Sheldon values 3, 4, 8, 12, 20, 40, 50, 55, 58, and 60 through 70. The continuous score e_sheldon is the probability-weighted expected Sheldon value, e_sheldon = sum_i p_i times grade_i. On the final coin-level probability vectors, the maximum absolute deviation of the probability sum from 1 was 2.52e-7. Probability of superiority, equivalent to normalized Mann-Whitney U, is invariant to any strictly monotonic transformation of the score. Pure rescaling, global calibration error, or compression and expansion of the Sheldon output therefore cannot by themselves create a within-grade service rank shift. Service-correlated nuisance signals remain possible and are addressed through source, side, identity, old/new, both-sides, and external-cohort checks.

4.5 Measurement-system validation and holder/service cues

Before grading, the coin is isolated from the source photograph and non-coin pixels are replaced with a neutral background. A post-analysis verification of 500 primary-corpus Morgan images sampled across sides and folds found that a mean 99.24% of pixels outside the actual production coin mask were neutral gray (minimum 97.21%), confirming that direct certification labels, service logos, inserts, barcodes, and surrounding slab geometry are ordinarily excluded from the grader input. This substantially reduces the conventional holder-recognition confound. It does not make service-correlated information mathematically impossible: small boundary fragments, rim-adjacent reflections, or image-acquisition characteristics could survive preprocessing. No correctly implemented v9 residual-cue ablation has yet quantified whether such remaining signals change numerical output. The empirical grade-specific pattern nevertheless constrains a simple service-cue explanation: a constant PCGS-versus-NGC visual boost would not readily produce strong upper-AU separation together with near-parity through MS-62 to MS-66. Residual grade-dependent interactions cannot be completely excluded.

4.6 Behavioral checks supporting use as a comparative measurement instrument

Because the grading engine is proprietary, this study evaluates observable behavior relevant to its use as a comparative ruler rather than disclosing a construction recipe. The checks below address out-of-sample independence, ordered response, limited repeat-image stability, robustness across coin sides and model generations, and behavior in a separately held-out cohort. They support use for this comparative population analysis; they do not establish absolute grading accuracy.

Behavioral checks supporting use of the grading instrument as a comparative ruler.

Property	Observed evidence in this study	What it supports / does not support
Out-of-sample independence	Five-fold out-of-fold scoring was assigned at the physical-coin level. A later identity audit found 148 physical coins spanning fold identities; those coins were excluded before the final analysis.	Substantially limits coin-level memorization. It does not make the broader professional grading labels independent of training.
Output sanity and ordinal response	The 20-class probability vectors summed to one to numerical precision (maximum absolute deviation 2.52e-7). Mean continuous response rose with stated grade, and the newer grader showed materially greater high-Mint-State resolution than the prior grader.	Shows a coherent ordered measurement scale and numerically well-formed output. It is not proof that the professional label is ground truth.
Repeat-image stability	In the opportunistic same-coin, same-side	Direct evidence that ordinary

Property	Observed evidence in this study	What it supports / does not support
	repeat set (63 coin-side units), median absolute score difference was 0.26 Sheldon points, mean 0.44, and 90th percentile 1.03.	re-imaging/model wobble is smaller than the observed same-label coin-to-coin spread on average. The repeat set is small, PCGS-only, and not a purpose-built repeatability experiment.
Independent coin-side robustness	The principal AU-55 and AU-58 service differences were positive when obverse and reverse predictions were analyzed separately. AU-55: d=0.232 / 0.181; AU-58: d=0.272 / 0.302 (obverse / reverse).	Reduces concern that the main population result is driven by one face of the coin. This is a robustness check, not a direct test of absolute accuracy.
Model-generation robustness	The prior and current graders were compared on essentially the same physical coins and images (99.88% exact image membership). Despite materially different coin-level judgments and improved high-MS resolution, both reproduced the principal upper-AU localization.	Shows that the main pattern is not peculiar to one numerical calibration. The two graders share development lineage and therefore are not independent external instruments.
External-cohort robustness	A separately held-out classic-Morgan cohort contained 27,272 physical coins and 50,213 image rows and was excluded before v9 training. All five frozen fold-model pairs reproduced positive AU-55/AU-58 separation; the commercial MS core remained much closer to parity.	Reduces concern that the result is peculiar to the primary CV-verified cohort. It does not establish objective grading correctness.
Holder/background mitigation	The production pipeline isolates the coin and neutralizes non-coin pixels before inference. In a 500-image verification sample, mean neutrality outside the actual production mask was 99.24% (minimum 97.21%).	Direct holder identifiers and most slab geometry are therefore ordinarily absent. Small residual boundary/reflection/acquisition cues remain possible; see Section 4.5.

These checks support the use of the system as a stable comparative measurement instrument; they do not establish absolute grading accuracy. Professional certification labels are not an independent physical reference standard because professional labels informed model development. Out-of-fold evaluation prevents the analyzed physical coin from being scored by a model trained on that same coin, but it does not remove dependence on the broader professional-label system. Accordingly, the study interprets the instrument as a common ruler for relative population comparisons, not as a definition of true grade.

A separate instrument-validation and repeatability report is in preparation for future release. That work is intended to use a purpose-built repeated-photography design, multiple independently acquired images per side, broader service representation, and variance-decomposition/repeatability statistics. The duplicate and repeated observations identified during corpus deduplication are also being preserved as a potentially useful supplementary source for that work. The expanded report is deliberately separated from the present Morgan service-comparison paper so that proprietary construction details can remain confidential while the instrument's observable stability is documented more completely.

5. Statistical Analysis

All confirmatory inference was performed at the physical-coin level. Repeated photographs within a side were averaged first; when both sides were present, the side means received equal weight. The primary service contrast was PCGS minus NGC within stated grade.

5.1 Effect measures

- Raw continuous-score gap: mean PCGS score minus mean NGC score within stated grade.
- Cohen's d: raw mean difference divided by pooled within-grade standard deviation [14].
- Probability of superiority (PoS): P(random PCGS score > random NGC score) + 0.5 times P(tie). A value of 50% indicates rank parity [12,13].

- Physical-coin nonparametric bootstrap confidence intervals. Direct grade-to-grade contrasts were recomputed inside each bootstrap replicate [15].

5.2 Frozen direct contrasts and classification

Before new-grader output review, strong replication was defined as both AU-55 and AU-58 remaining positive, satisfying d at least 0.20 and PoS at least 55%, exceeding AU-50 on direct d and PoS contrasts, and producing a larger mean upper-AU effect than the MS-63 through MS-66 commercial core. A grade met the practical-parity criterion only if its entire 95% confidence interval for d fell inside [-0.10,+0.10].

5.3 Composition and source controls

Date/mint control compared PCGS and NGC only within identical year, mint, and stated-grade strata. Bucket-count summaries used the predeclared minimum of 30 physical coins per service. The registered controlled PoS estimator aggregated Mann-Whitney pair counts across strata, theta = sum(U_h) / sum(n_P,h times n_N,h), rather than averaging stratum PoS values equally.

Source controls included within-source PoS, source+year+mint+grade pair-count-weighted PoS, and leave-one-source-out analysis over all 40 confidential sources. Source identities are not disclosed.

5.4 Instrument and cohort robustness

The analysis compared the old and new graders on the exact common physical-coin set. An image-membership audit showed that 99.88% of common coins used exactly the same image rows in both graders, 0.10% used the same side coverage with different individual images, and 0.02% differed in side coverage. Additional robustness analyses examined obverse and reverse separately, restricted identity to cert-anchored coins, restricted the primary analysis to physical coins with both sides present, and scored the fully held-out external CV-gate cohort with all five frozen fold-model pairs. The frozen specification also described a cross-side model-uncertainty hygiene analysis, but it did not uniquely specify the uncertainty subset rule, eligible both-side cohort, or service-separation measure. Because executing it after result review would require a post hoc analytic choice, it was explicitly reported as not completed rather than improvised.

5.5 Separately frozen exploratory follow-up

After the confirmatory study was closed, a separate exploratory specification was frozen and hashed before any exploratory result was inspected (SHA-256 9e3c2f2f55398dace90787a4574a2da98614312df9deafbcb891dd82e6484820). Two analyses were defined on the final classic-Morgan clean primary cohort. First, internal dispersion was compared within year×mint×stated-grade cells containing at least 30 PCGS and 30 NGC physical coins using SD, IQR, MAD, cell-level bootstrap intervals, and equal-cell aggregate summaries. Second, all eligible issue cells were screened for service-population outliers using raw gap, Cohen d, and PoS relative to a leave-that-issue-out baseline for the same stated grade.

For issue-outlier screening, excess PoS was the primary ranking statistic. Bootstrap sign p-values were adjusted across all eligible cells using the Benjamini-Hochberg false-discovery-rate procedure [16]. Candidate robustness checks were restricted to the prespecified candidate set and examined obverse/reverse direction, old-versus-new grader direction, and confidential source concentration. These analyses are explicitly exploratory. An extreme cell may be consistent with what collectors sometimes describe as a guarded grade, but the design cannot establish deliberate grade protection, human-grader policy, or objective grading correctness.

5.6 Absolute same-label dispersion follow-up

After the consistency/outlier follow-up was closed, a second descriptive exploratory specification was frozen before result inspection (SHA-256 7a420ab655b1cea46ec73f1261de5a6b7b7485e1ec2a485a435cfdc74fbb61e4). It reused the same 207 eligible year x mint x stated-grade cells, each containing at least 30 PCGS and 30 NGC physical coins. For each service-specific cell, the analysis reported absolute model-score SD, IQR, MAD, score quantiles, and

residuals relative to the stated label. Residuals were explicitly defined as measurement-system disagreement, not grading error.

The analysis also calculated the distribution of absolute score differences between two randomly selected physical coins from the same year, mint, stated grade, and service, including probabilities of differences of at least 0.5, 1.0, 1.5, and 2.0 Sheldon points. A diagnostic repeat-image analysis measured same-coin, same-side score variation where repeated independently scored images were available. This opportunistic repeat set is the study's only direct empirical check on short-run measurement repeatability; it was used to judge whether observed same-label spread could plausibly be explained entirely by re-imaging/model wobble, not to subtract a measurement-error variance component. Raw Sheldon-point dispersion across circulated grades is not equivalent to a count of grade steps because the numerical spacing between adjacent circulated labels is much larger than the one-point spacing of most Mint State labels.

5.7 Date/mint average-stability follow-up

After the absolute-dispersion follow-up was closed, a final narrowly scoped exploratory specification was frozen before result inspection (SHA-256 680ca792a702930dadc0d26524035986679c992abaf7a8ae8def121bb479f774). It reused exactly the same 207 eligible year x mint x stated-grade cells and the same v9 physical-coin scores. Within each service and stated grade, the analysis estimated the mean score for every eligible date/mint issue (the issue's average score, termed the 'issue center' in the analysis) and compared that average with an equal-issue-weight grade baseline. The resulting issue shift therefore measures whether a date/mint's average score sits visually above or below the typical eligible date/mint carrying that same service and label under the common ruler.

For each grade and service, the analysis summarized the range, SD, MAD, and quantiles of issue-cell means. PCGS and NGC shifts for the same year/mint/grade were then paired to form a common issue shift, the mean of the two service-specific shifts, and a service-differential shift. Concordance was summarized with Pearson and Spearman correlations and sign agreement. This analysis is descriptive and does not identify rarity, strike, luster, die state, market sorting, or an issue-relative grading policy as the cause of a shift. Because the grader learned from professional labels and can see issue-identifying information on the coin, it is not an issue-blind test of an abstract grading standard.

6. Frozen Confirmatory Hypotheses

1. PCGS-NGC separation is materially larger at AU-55 and AU-58 than at AU-50.
2. Upper-AU separation is materially larger than separation through the commercial Mint State core, particularly MS-63 through MS-66.
3. Separation falls sharply after AU-58 as stated grade progresses through MS-60, MS-61, and MS-62. Strict monotonic decline is not required.
4. Any renewed difference at MS-67 or MS-68 is evaluated separately from the upper-AU hypothesis.
5. No causal mechanism is prespecified.

Confirmatory versus exploratory boundary
The old grader generated the hypothesis. The new grader and the analyses described above provide the confirmatory replication. ANACS, ICG, individual date/mint outliers, grade-guarding claims, price-boundary analyses, and same-coin crossover studies remain exploratory or future work.

7. Results

7.1 Primary grade-by-grade comparison

Grade	n PCGS	n NGC	Raw gap	Cohen d	PoS	Interpretation
AU-50	1,355	1,253	+0.372	+0.079	52.1%	Small positive effect
AU-55	738	768	+0.731	+0.227	56.4%	Strong upper-AU separation
AU-58	792	824	+0.790	+0.304	58.7%	Strongest primary separation
MS-60	277	316	+0.385	+0.115	53.6%	Modest; interval includes small effects
MS-61	468	734	+0.170	+0.100	53.3%	Modest; interval includes small effects
MS-62	1,829	1,740	-0.019	-0.017	49.7%	Practical parity
MS-63	5,685	4,400	-0.009	-0.012	50.0%	Practical parity
MS-64	9,655	6,686	+0.009	+0.014	50.5%	Practical parity
MS-65	7,745	5,201	+0.009	+0.012	50.9%	Practical parity
MS-66	3,581	2,239	-0.013	-0.024	48.8%	Practical parity
MS-67	555	794	+0.073	+0.119	52.9%	Small upper-MS return
MS-68	6	16	-0.349	-0.454	39.6%	Descriptive only; inadequate n

AU-55 and AU-58 both satisfied the predeclared strong-magnitude reference. AU-50 was positive but substantially smaller. MS-62 through MS-66 met the preregistered practical-parity criterion because the entire bootstrap confidence interval for d fell inside [-0.10,+0.10]. MS-67 showed a smaller positive effect with d=0.119 and a d interval entirely above zero; MS-68 was too sparse for inference. The final registered classic-date scope correction removed one NGC MS-65 coin and one PCGS MS-66 coin from the headline range; all resulting changes in raw gap, d, and PoS were 0.001 or smaller, and every qualitative conclusion was unchanged.

Figure 1. Primary confirmatory Cohen's d by stated grade. The shaded band marks the preregistered practical-parity reference of -0.10 to +0.10.

Figure 2. Scale-free probability of superiority by stated grade. Fifty percent represents rank parity; the dotted line marks the predeclared 55% strong-magnitude reference.

7.2 Direct confirmatory contrasts

Contrast	Measure	Point estimate	95% bootstrap CI
AU-55 vs AU-50	d	+0.148	[+0.014,+0.277]
AU-55 vs AU-50	PoS	+4.3 points	[+0.5,+8.0]
AU-58 vs AU-50	d	+0.226	[+0.106,+0.362]
AU-58 vs AU-50	PoS	+6.6 points	[+3.2,+10.4]
Mean AU-55/58 vs mean MS-63/66	d	+0.268	[+0.191,+0.342]
Mean AU-55/58 vs mean MS-63/66	PoS	+7.5 points	[+5.4,+9.5]
AU-58 vs MS-60	d	+0.189	[+0.026,+0.373]
AU-58 vs MS-60	PoS	+5.1 points	[-0.0,+10.5]
AU-58 vs MS-61	d	+0.204	[+0.053,+0.365]
AU-58 vs MS-61	PoS	+5.4 points	[+1.4,+9.5]
AU-58 vs MS-62	d	+0.322	[+0.205,+0.440]
AU-58 vs MS-62	PoS	+9.0 points	[+5.6,+12.3]

The direct standardized and rank contrasts satisfy the frozen strong-replication definition. The final classic-scope corrective output supplied the previously missing AU-58 versus MS-61 d contrast: +0.204 with a 95% bootstrap CI of [+0.053,+0.365]. The scope correction did not change any AU-50 through MS-62 point estimate; the mean upper-AU versus MS-63 through MS-66 d contrast changed only from +0.26794 to +0.26773 because one modern coin was removed from MS-65 and one from MS-66. The table reports rounded values from the final classic-scope corrective output; full-precision endpoints are retained in the frozen CSV and manifest.

7.3 Date/mint and source controls

Grade	Year+mint controlled PoS	Source+year+mint controlled PoS
AU-50	53.0% [49.9,56.2]	53.1% [50.1,56.0]
AU-55	51.3% [46.6,55.8]	53.5% [49.8,57.4]
AU-58	54.9% [50.7,59.1]	55.7% [52.1,59.5]
MS-60	No eligible n>=30 buckets	56.2% [50.4,62.1]
MS-61	64.2% [55.3,73.1]; 2 buckets	55.7% [51.6,60.1]
MS-62	51.3% [48.3,54.5]	52.0% [49.5,54.5]
MS-63	52.4% [50.9,53.8]	52.6% [51.2,54.1]
MS-64	52.7% [51.6,53.7]	52.9% [51.9,54.0]
MS-65	52.1% [51.0,53.3]	52.2% [51.0,53.4]
MS-66	49.3% [47.5,51.1]	49.2% [47.6,50.9]
MS-67	50.8% [46.5,55.0]	51.3% [47.7,55.1]

AU-58 remained positive under same-year/mint, within-source, and fully source+year+mint controls. AU-55 remained strong within source and under every leave-one-source-out run, but it was more heterogeneous across large

year/mint cells: only three of six eligible n>=30/n>=30 cells were positive, and the pair-weighted year/mint PoS interval crossed 50%. The table reports the two composition-matched estimators with complete frozen intervals; the full within-source table is retained in the reproducibility output package. The primary AU-55 effect replicated strongly, but its issue-level breadth was weaker than AU-58.

The source controls did not support a single-source explanation for the overall localization. Across all 40 leave-one-source-out runs, the mean upper-AU d and centered-PoS effects exceeded the MS-63 through MS-66 mean. The smallest localization margins occurred when the dominant Morgan archive source was removed, but neither d nor PoS reversed. Fully source+year+mint-stratified upper-AU estimates were based largely on the source with sufficient cross-service overlap, so they should not be described as broad multi-source replication on their own.

Conditional estimates also revealed small positive PCGS advantages at MS-63 through MS-65 even though the primary all-population estimates met practical parity. These are different estimands. The primary result describes the overall current holder populations; the stratified estimates compare services within matched issue/source cells and can expose small within-cell effects that are offset by population composition.

7.4 Old-versus-new grader replication

All 65,092 final classic-scope primary physical coins had a usable v8 comparator. In the pre-scope image-membership audit, image membership was exactly identical for 99.88% of coins, the same side coverage with different images for 0.10%, and different side coverage for 0.02%. The classic-scope correction removed 91 modern physical coins; only two of them entered MS-65 or MS-66 in the headline range, so the old-versus-new grade-shape conclusions were unchanged.

Measure	v8 old grader	v9 new grader	Interpretation
AU-50 raw gap	+0.396	+0.372	Essentially unchanged
AU-55 raw gap	+1.012	+0.731	Smaller; change interval includes zero
AU-58 raw gap	+0.622	+0.790	Larger; change interval includes zero
MS-63 raw gap	+0.121	-0.009	Old small effect erased
MS-64 raw gap	+0.179	+0.009	Old small effect erased
MS-65 raw gap	+0.155	+0.009	Old small effect erased
MS-66 raw gap	+0.158	-0.013	Old small effect erased/reversed
MS-67 raw gap	+0.074	+0.073	Essentially unchanged
AU-50 to AU-58 grade-response slope	0.813	1.095	Greater AU resolution
MS-62 to MS-67 grade-response slope	0.536	0.704	Substantially greater MS resolution

Within-grade old/new score agreement was moderate in upper AU and low through much of Mint State. Pearson r was 0.591 at AU-55, 0.468 at AU-58, 0.213 at MS-63, 0.176 at MS-64, and 0.204 at MS-65. The new model was therefore not merely a monotonic rescaling of the old one. It made substantially different coin-level judgments and had much greater high-MS resolution, yet the upper-AU population localization replicated while the old grader's small MS-63 through MS-66 service differences largely disappeared.

7.5 Obverse and reverse robustness

Grade	Side	Cohen d	PoS
AU-50	Obverse	+0.073	51.8%
AU-50	Reverse	+0.088	51.9%
AU-55	Obverse	+0.232	56.9%
AU-55	Reverse	+0.181	55.2%
AU-58	Obverse	+0.272	57.7%
AU-58	Reverse	+0.302	58.6%
MS-60	Obverse	+0.071	52.9%
MS-60	Reverse	+0.196	56.9%
MS-61	Obverse	+0.052	51.7%
MS-61	Reverse	+0.119	54.3%
MS-62 through MS-66	Both faces	Near zero	Approximately parity
MS-67	Obverse	+0.072	51.2%
MS-67	Reverse	+0.045	50.6%

AU-55 and AU-58 were positive independently on both faces. The result is therefore not an obverse-only artifact. The low-MS residual was more visible on the reverse at MS-60 and MS-61, but the analysis did not prespecify or establish a side-specific causal mechanism.

7.6 Cert-anchored identity sensitivity

Restricting the analysis to service-consistent certification-anchored physical identities left the headline result essentially unchanged. AU-55 remained at d=0.227 and PoS=56.4%; AU-58 remained at d=0.305 and PoS=58.5%. MS-62 through MS-66 remained near zero, and MS-67 remained a small positive result. AU-50 weakened to d=0.033 and its gap interval crossed zero. The cert-only table was re-filtered under the final classic-date scope; the two modern MS-65/MS-66 headline coins were removed without changing any qualitative conclusion.

7.7 External CV-gate sensitivity cohort

Grade	n PCGS	n NGC	Ensemble d	Ensemble PoS
AU-50	362	282	+0.159	54.2%
AU-55	179	177	+0.395	61.4%
AU-58	328	293	+0.459	64.1%
MS-60	135	152	+0.542	67.3%
MS-61	229	256	+0.221	56.0%
MS-62	784	545	+0.000	50.2%
MS-63	1,625	941	-0.055	48.7%

Grade	n PCGS	n NGC	Ensemble d	Ensemble PoS
MS-64	2,499	1,329	-0.062	48.5%
MS-65	2,502	1,217	+0.012	50.7%
MS-66	2,062	858	-0.036	49.1%
MS-67	1,064	798	+0.198	56% (rounded)

The fully held-out classic external cohort reproduced the upper-AU separation and the MS-62 through MS-66 near-zero band. After the final scope correction it contained 27,272 physical coins and 50,213 image rows, with zero physical-coin overlap with v9 train/validation/test. Every one of the five separately trained fold-model pairs produced positive AU-55 and AU-58 effects. The external cohort also produced a large, consistent MS-60 effect, with d between 0.45 and 0.56 across all five fold pairs. A post-hoc physical-coin bootstrap on the frozen external ensemble gave MS-60 d=0.542 (95% CI [0.309, 0.789]), compared with primary d=0.115 (95% CI [-0.038, 0.246]); the intervals do not overlap. This is a substantive cohort difference rather than ordinary sampling overlap.

MS-60 is not the only grade at which the external estimate is larger, but it is the largest cohort divergence. A post-hoc source-composition diagnostic found that both primary and external MS-60 populations were dominated by the same anonymous dealer source. Restricting to common sources left the discrepancy essentially unchanged, and within the dominant source d was approximately 0.075 in the primary cohort versus 0.505 in the external cohort. Source composition therefore does not explain the difference. Probability-vector diagnostics further showed that external NGC MS-60 was broadly similar to primary NGC MS-60, whereas the external PCGS MS-60 population shifted upward across the score distribution rather than through a single extreme tail. The design cannot determine why those PCGS populations differ, so the cohort discrepancy remains open rather than being attributed to grading policy, verification selection, or model artifact.

The frozen external headline output did not include bootstrap confidence intervals; the intervals reported above were added later as a post-hoc uncertainty characterization using the already frozen external ensemble and physical-coin unit. The external results remain sensitivity analyses. Importantly, the preregistered immediate-decay clause did not replicate in this cohort: rather than falling immediately after AU-58, the external effect rose to its maximum at MS-60 and then fell through MS-61 to essentially zero at MS-62. Thus the narrow claim that separation peaks at AU-58 and immediately decays is cohort-dependent.

Figure 3. Primary OOF and external held-out sensitivity cohorts. Both cohorts reproduce upper-AU separation and MS-62 through MS-66 near-parity. External effects are generally larger at grades with nonzero separation, most dramatically at MS-60, where the external cohort peaks rather than showing the primary cohort's immediate post-AU-58 decline.

7.8 Probability-vector diagnostics

The probability-vector diagnostic was recomputed at the registered physical-coin unit. The model uses 20 numerical grade classes and e_sheldon is the probability-weighted expected Sheldon value; coin-level probability sums were numerically valid, with maximum absolute deviation from 1 of 2.52e-7. At AU-55, PCGS coins assigned 34.52% mean probability mass to AU-58 or higher versus 30.55% for NGC, a +3.97 percentage-point shift. At AU-58, PCGS coins assigned 41.16% mean mass to MS-60 or higher versus 33.37% for NGC, a +7.79-point shift. The full mean vectors show a broad redistribution toward higher neighboring grades rather than a conclusion about objective correctness. These diagnostics describe where the model distribution moves; they do not identify a causal grading mechanism.

7.9 Remaining preregistered robustness outputs

The separately preregistered both-sides-present robustness analysis was completed using equal-weight obverse/reverse coin scores. Among PCGS/NGC coins in AU-50 through MS-67, 49,388 physical coins had both sides represented. AU-55 remained d=0.230, PoS=56.5%, and AU-58 remained d=0.329, PoS=58.9%; MS-62 through MS-66 remained inside the practical-near-parity band, and MS-67 remained a smaller positive effect. The final classic-scope correction affected only one MS-65 and one MS-66 headline coin and did not alter this conclusion. The predeclared cross-side uncertainty analysis was not completed: the frozen plan did not uniquely specify the uncertainty subset rule, both-side eligibility definition, or service-separation measure, so executing it after result review would require a post hoc analytic choice. It is therefore reported transparently as uncompleted secondary robustness rather than improvised.

7.10 Exploratory follow-up: internal consistency and issue-level heterogeneity

The exploratory consistency analysis included 207 year×mint×grade cells with at least 30 physical coins from each service. PCGS had the smaller SD in 49.3% of cells, the smaller IQR in 54.1%, and the smaller MAD in 52.2%. The overall median SD ratio was 1.001, while the equal-cell geometric mean SD ratio was 0.968 (95% CI 0.938-0.995). This is a small aggregate difference rather than evidence that PCGS is uniformly more consistent. The pattern was grade-dependent: MS-66 showed the clearest tighter-PCGS dispersion (80% of 20 cells with smaller PCGS SD; geometric mean SD ratio 0.870, 95% CI 0.784-0.938), whereas several other grades were near parity or modestly favored tighter NGC dispersion.

Grade / summary	n cells	% PCGS SD < NGC	Median SD ratio	Geo-mean SD ratio [95% CI]
OVERALL	207	49.3%	1.001	0.968 [0.938, 0.995]
VG-8	6	66.7%	0.671	0.726 [0.544, 0.980]
VF-20	6	83.3%	0.942	0.961 [0.888, 1.054]
EF-40	6	16.7%	1.065	1.037 [0.930, 1.128]
AU-50	14	28.6%	1.027	0.990 [0.928, 1.048]
AU-55	6	33.3%	1.012	1.051 [0.960, 1.177]
AU-58	7	57.1%	0.979	0.949 [0.848, 1.042]
MS-62	14	42.9%	1.108	1.094 [0.955, 1.262]
MS-63	40	32.5%	1.069	1.044 [0.980, 1.106]
MS-64	44	50.0%	0.998	0.969 [0.921, 1.016]
MS-65	38	55.3%	0.953	0.936 [0.869, 0.996]
MS-66	20	80.0%	0.889	0.870 [0.784, 0.938]

The issue-level search found substantial heterogeneity but no statistically confirmed outlier after multiplicity control. None of the 207 eligible cells reached q_excess_PoS<0.05; the minimum q was 0.083 for 1896-P MS-66. The prespecified candidate table therefore consists of the largest positive and negative excess-PoS cells rather than discoveries passing FDR. The largest positive candidate was 1882-CC MS-64 (excess PoS +0.190), but the old grader pointed in the opposite direction, weakening its replication. More internally coherent examples included 1878-P MS-61 and 1880-CC MS-65 in the PCGS-favoring direction, and 1887-O MS-62, 1902-S MS-63, and 1896-P MS-66 in the NGC-favoring direction.

Issue	Grade	n P/N	d	PoS	Excess PoS	BH q	Side / old-new
1882-CC	MS-64	51/34	+0.684	0.694	+0.190	0.116	same / disagree
1878-P	MS-61	47/36	+0.597	0.703	+0.183	0.116	same / agree
1887-O	MS-62	42/40	-0.609	0.330	-0.171	0.237	same / agree
1892-P	AU-55	36/41	-0.341	0.414	-0.159	0.406	same / agree

Issue	Grade	n P/N	d	PoS	Excess PoS	BH q	Side / old-new
1894-O	AU-58	54/50	-0.165	0.450	-0.146	0.339	same / disagree
1902-S	MS-63	45/32	-0.542	0.356	-0.145	0.419	same / agree
1893-CC	EF-40	36/30	-0.681	0.311	-0.139	0.476	same / agree
1880-CC	MS-65	160/62	+0.541	0.644	+0.138	0.116	same / agree
1880-CC	MS-64	53/31	+0.498	0.642	+0.138	0.419	same / agree
1896-P	MS-66	238/96	-0.457	0.360	-0.137	0.083	same / agree

Robustness checks narrowed the interpretation further. Eighteen of 20 prespecified candidates had the same direction on obverse and reverse, and 15 of 20 had the same direction under the old and new graders. However, every candidate cell was highly concentrated in one confidential source, typically with the largest source supplying at least about 90% of both service populations. Within-source direction matched the cell-level direction, but this concentration prevents strong generalization from a candidate cell to a service-wide issue policy. The appropriate description is an issue-specific service-population outlier, not evidence of deliberate grade guarding.

7.11 Exploratory follow-up: absolute same-label visual-quality dispersion

Absolute within-label dispersion was substantial enough to be practically visible but was similar between services. Across the 207 eligible issue cells, the equal-cell-weighted median SD of coin-level model scores was 0.63 Sheldon points for PCGS and 0.65 for NGC. Observation-weighted SD was approximately 1.42 and 1.45, respectively, because lower circulated and low-AU cells have much larger raw Sheldon spacing and therefore dominate an observation-weighted raw-point summary. The service comparison itself remained close: the analysis did not reveal a broad absolute-dispersion advantage for either PCGS or NGC.

Service	Median cell SD	Median P(|pair diff| >= 1.0)	Median P(|pair diff| >= 2.0)	Observation-weighted SD
PCGS	0.63	0.24	0.033	~1.42
NGC	0.65	0.26	0.034	~1.45

The pairwise quantities provide the most intuitive interpretation. In the median eligible cell, two randomly selected coins carrying the same year, mint, stated grade, and service label differed by at least 1.0 modeled Sheldon point about 24% of the time for PCGS and 26% for NGC. Differences of at least 2.0 points occurred in about 3.3% and 3.4% of median cells. In MS-63 through MS-66, where one Sheldon point is approximately one adjacent numeric grade step, the probability of a >=1-point pair difference ranged roughly 13% to 27%, while >=2-point differences were about 0.4% to 4%. At VF-20 through AU-50, raw-point dispersion was much larger, but those raw numbers should not be read as counts of adjacent grade steps because the Sheldon labels are widely spaced there.

The repeat-image diagnostic identified 63 same-coin, same-side repeat-image units, all within the PCGS population. Their median absolute score difference was 0.26 Sheldon points, mean 0.44, and 90th percentile 1.03. Mid-Mint-State same-label pairs had median pair differences of approximately 0.60-0.74 points, materially larger than the repeat-image median. The diagnostic therefore argues against explaining all within-label dispersion as repeat-image noise, but the repeat sample is too small and unrepresentative for a formal measurement-error decomposition.

7.12 Exploratory follow-up: does a stated grade have the same average quality across date/mints?

The date/mint-average analysis found that date/mint changes the center of a same-labeled population much less in commercial Mint State than the raw circulated-grade Sheldon ranges initially suggest. In MS-63, eligible issue means spanned 1.21 points for PCGS and 0.91 for NGC, with SDs of issue means of 0.26 and 0.22. At MS-64 the ranges were 0.78 and 0.75, with SDs 0.19 and 0.18; at MS-65, 0.74 and 0.71 across 38 eligible issues, with SD 0.16 for both services; and at MS-66, 0.53 and 0.56, with SD 0.18 for both. Thus the extreme issue-to-issue range can approach roughly one Mint State grade step at MS-63, but the typical variation of issue centers is only a fraction of a grade

step. This between-issue variation is smaller than the approximately 0.5-0.7 point within-label SD observed for individual coins in the same commercial Mint State range.

The two services tended to move the same issues in the same direction. Across all 207 cells after centering within stated grade, the explicit all-cell result was Pearson r=0.657, with 82.1% sign agreement. Concordance was strongest in mid/high Mint State, where grade-specific Pearson correlations were approximately 0.81-0.85 and roughly 76-90% of eligible cells shared direction. This common movement is compatible with real issue-wide production characteristics, industry-wide issue-relative conventions, or common measurement-system inheritance; the present design cannot separate these explanations.

Only two 1921 issue/grade cells met the n>=30-per-service eligibility requirement: 1921-P MS-64 and MS-65. Their common issue shifts were small and negative, -0.16 and -0.24 Sheldon points, respectively. Those values are small relative to the roughly 0.55-0.65 point within-cell SD in those grades. The eligible 1921-P results therefore do not show a large downward center shift under v9. They do not test 1921-S, lower grades, or strike directly, and a null result under v9 cannot exclude an industry-wide issue allowance that the model itself learned.

The largest positive and negative raw common shifts occurred in circulated and low-AU cells rather than MS-63 through MS-66. Because circulated Sheldon labels are unevenly spaced, those raw ranges should not be interpreted as a larger number of adjacent grade steps without local normalization. No commercial MS-63 through MS-66 cell appeared among the ten largest positive or ten largest negative raw common shifts.

8. Discussion

8.1 Principal finding

The preregistered grade-specific hypothesis met the registered Strong Replication classification in the primary confirmatory cohort. PCGS and NGC do not differ by a uniform offset across the Morgan grading scale. In the primary population, the clearest separation occurs at AU-55 and AU-58, is smaller at AU-50, decays through MS-60 and MS-61, and reaches practical parity from MS-62 through MS-66. A smaller positive PCGS effect returns at MS-67.

The external cohort reproduced the upper-AU separation and commercial-Mint-State convergence but did not reproduce the registered immediate decline after AU-58. Its largest effect occurred at MS-60 (ensemble d about 0.54), after which separation fell through MS-61 to essentially zero at MS-62. Effect magnitudes were also generally larger in the external cohort at the other non-parity grades, so MS-60 should not be treated as an isolated anomaly. Across both cohorts, the more robust description is a non-uniform service-population difference concentrated around the upper-AU/low-Mint-State boundary, followed by convergence through MS-62 to MS-66 and a smaller upper-MS return. The narrower 'AU-58 peak followed by immediate decay' description is supported by the primary cohort but is not stable across cohorts.

8.2 Why the result is difficult to dismiss as a generic model artifact

- The hypothesis specified a localized shape rather than merely a positive PCGS coefficient.
- The scale-free PoS result is not created by monotonic compression or expansion of the Sheldon output.
- As detailed in Section 4.5, direct holder identifiers are removed before inference. Small residual cues cannot be excluded, but a simple constant service-recognition boost is inconsistent with the strongly grade-specific pattern.
- The improved grader made substantially different coin-level judgments and had much greater high-MS resolution, yet reproduced the upper-AU localization on essentially identical images.
- AU-55 and AU-58 were positive on both obverse and reverse, and the separate both-sides-present restriction reproduced the same localization.
- Restricting to cert-anchored physical identity left the finding unchanged.
- The localization remained positive under all 40 leave-one-source-out runs.
- A fully held-out, CV-unverified external population reproduced upper-AU separation across all five frozen fold-model pairs.

8.3 What the result means for the market

The scientific object is the current holder population. A buyer choosing a same-date, same-mint PCGS or NGC coin at a fixed label draws from the population that exists after grading, selective submission, crackouts, crossovers, reconsiderations, resubmissions, and market sorting. The study therefore answers a practical population question even though it cannot causally attribute the difference to the original grading desk.

8.4 Mechanisms remain open

Several mechanisms can coexist. PCGS may use different effective thresholds at specific boundaries. Dealers may selectively submit stronger or weaker candidates to different services. High-end NGC coins may migrate into PCGS holders through crossovers or crackouts. Historical grading standards and local human-grader effects may differ. The observed grade localization is coherent with a boundary process, but the present design does not identify the causal decomposition. The paper should not state that wear, luster, price cliffs, protected grades, or deliberate behavior has been proven.

8.5 Heterogeneity matters

AU-58 was more robust than AU-55 under large-cell date/mint matching. AU-55 remained strong overall and within the dominant source but showed mixed signs across the six largest eligible issue cells. The separately frozen exploratory search confirmed substantial issue-level heterogeneity in both directions, but no cell survived BH FDR at $q<0.05$. The strongest candidates should therefore be treated as replication targets rather than established guarded grades.

8.6 Exploratory implications: consistency and possible guarded-grade patterns

The dispersion analysis does not support a simple claim that one service is uniformly more consistent. Overall PCGS and NGC within-cell dispersion was very similar, and the fraction of cells with smaller PCGS SD was essentially 50/50. The clearest exception was MS-66, where PCGS holder populations were materially tighter across the eligible issue cells. This could reflect grading thresholds, market sorting, submission behavior, or other population processes; it should not be interpreted as direct evidence about individual grader repeatability.

The issue-level analysis provides a disciplined way to investigate collector claims about grades that a service "guards." Several cells show large departures from their same-grade baseline, including both PCGS-favoring and NGC-favoring directions. Yet none survived the prespecified false-discovery threshold, some failed old/new replication, and all candidates were dominated by one source. These results justify targeted independent replication, ideally from a second dealer or archive population, but they do not establish intentional grade protection.

8.7 Same label does not imply visual interchangeability

The absolute-dispersion follow-up makes a different point from the service comparison: coins carrying the same date, mint, service, and stated grade remain meaningfully heterogeneous under the common visual ruler. In the commercial Mint State range, where adjacent numeric grades are separated by one Sheldon point, random same-label pairs differ by at least one modeled point often enough to be practically relevant. The date/mint-average analysis sharpens this result: in MS-63 through MS-66, the SD across date/mint average scores is only about 0.16-0.26 points, while the SD among individual coins within an exact date/mint/service/grade group is roughly 0.5-0.7 points. In other words, coins within one exact date/mint/service/grade group can differ substantially from one another even though the average score for that grade changes relatively little from one well-populated date/mint to another. These are not estimates of grading error, and the model is not an objective answer key. They nevertheless provide unusually direct quantitative support for “buy the coin, not the holder”: the holder grade strongly locates a quality band, but it does not make individual coins visually interchangeable.

Figure 4. Two kinds of variation within commercial Mint State (MS-63 through MS-66). The upper range shows the typical standard deviation among individual coins that share the same date, mint, grading service, and stated grade. The lower range shows the standard deviation of the average score from one eligible date/mint to another at the same stated grade. Thus, individual coins within a single labeled group vary substantially more than the average quality level changes across date/mints. Values are instrument-relative population dispersion, not grading error.

A related future-study opportunity comes from the deduplication work itself. Repeated appearances judged to represent the same certification or physical coin were intentionally collapsed or removed to protect independence in the present population analysis. Those linked repeat observations should be preserved rather than discarded as mere pipeline debris. With a high-confidence visual-fingerprint linkage protocol, they could support a separate study of same-coin photographic stability, repeated sales, reholdering, and possible crossover or crackout trajectories. Such a study would answer a different question from the present holder-population comparison and should remain separate.

8.8 How much of the same-label spread could be measurement wobble?

The measuring instrument is not perfectly repeatable, and the same-label dispersion should not be interpreted as if every observed difference were a real difference between coins. The direct evidence available here is limited but informative. Among 63 same-coin, same-side units with independently scored repeat images, all from PCGS holders, the median absolute score difference was 0.26 Sheldon points, the mean was 0.44, and the 90th percentile was 1.03. By comparison, median pairwise differences between two distinct same-label coins in MS-63 through MS-66 were approximately 0.60-0.74 points. The repeat-image result therefore makes it difficult to explain all of the observed same-label spread as ordinary re-imaging or model wobble.

That diagnostic does not establish how much of the spread is “real coin quality.” The repeat sample is small, PCGS-only, opportunistic rather than experimentally designed, and limited to same-side repeats. It also does not test systematic measurement bias. Random image/model noise tends to cancel when many coins are averaged, so the comparatively stable date/mint averages are less sensitive to ordinary random wobble than any one coin score. Systematic issue-conditioned bias, learned grading conventions, or consistent image-acquisition effects would not necessarily cancel with larger samples.

A purpose-built repeatability study is being developed for separate future release. The planned objective is to score multiple independently captured images of the same physical coins and sides across controlled and ordinary capture conditions, then estimate repeatability directly, including a variance decomposition or comparable repeatability statistic. The high-confidence same-coin links preserved from the deduplication pipeline may provide an additional natural-repeat sample, but they should supplement rather than replace a deliberately captured repeatability set. Until that work is complete, the present paper treats measurement wobble as a quantified limitation rather than assuming it is zero.

8.9 Issue-specific standards are constrained by the data; strike remains open

The final date/mint average follow-up now answers part of the earlier question. Under the current v9 common ruler, a nominal commercial Mint State grade does not move dramatically from issue to issue among the well-populated date/mint cells studied. MS-65 is the clearest example: across 38 eligible issues, the range of issue means was 0.74 Sheldon points in PCGS and 0.71 in NGC, while the SD of issue means was only 0.16 in each service. MS-64 and MS-66 showed similarly small date/mint-average SDs. The data therefore do not support a simple claim that a PCGS or NGC MS-65 routinely means a substantially different visible quality level for each date/mint. Rare or key dates that fail the n>=30-per-service threshold remain outside this inference.

Published grading standards make the remaining standards question especially important. PCGS states that its written numeric-grade descriptions apply to all coin types and strikes and describes MS-65 as requiring an above-average strike; NGC publishes a generic numeric scale that describes MS-65 as well struck. [1,2] If a future issue-blind ruler were to show substantial same-grade quality shifts by date/mint, the nominal standard would not be visually invariant across issues even if issue-relative allowances are accepted market practice. The present v9 results instead show comparatively stable commercial Mint State centers, but they cannot prove that the official standard is applied issue-blind because v9 learned from professional grading labels and may inherit the same issue-conditioned conventions.

Morgan-dollar reference commentary hosted by PCGS illustrates why strike remains a live question. PCGS CoinFacts describes 1921 Morgans as characteristically poorly struck and notes that certified MS-65 examples can look markedly inferior to earlier-date MS-65 Morgans; its 1921-S commentary similarly describes grading interpretations as looser than for earlier dates. [3,4] This is expert reference commentary hosted by PCGS, not a formal PCGS policy statement. In the current data, only 1921-P MS-64 and MS-65 met the 30/30 eligibility threshold, and both were only slightly below their same-grade issue baselines (-0.16 and -0.24 common shifts). That does not support a large visible 1921-P center shift under v9, but it also does not falsify a weak-strike allowance because strike was not measured and the model may have learned the same convention.

The distinction between grade scarcity and actual quality scarcity is only partly constrained. The stable centers of well-populated MS-64 through MS-66 issues argue against large, routine issue-specific shifts in the visible meaning of those labels. But scarcity itself was not modeled, and many genuinely rare or key date/grade combinations will fail the 30/30 cell threshold. A rare MS-65 population could still be scarce because few coins possess the relevant quality, because the effective label threshold differs, or because submission, crossover, crackout, and market history alter the holder population. Rarity by itself should not be treated as the mechanism.

Weak strike cannot be isolated in the current corpus because strike quality was not independently annotated or measured. Date/mint differences in strike, luster, die state, and typical surface preservation therefore remain unmeasured issue-level factors. This is the key standards limitation: the published standards are presented generically, but the present model cannot tell whether an apparent standard is truly issue-independent or simply reproduces issue-relative professional practice. If strike or issue identity changes the grade threshold while the formal label is presented as generic, that is precisely where “standard” may cease to mean visually invariant standard. The current study cannot measure that directly.

A separate measurement-system stress test should target the remaining service-cue concern directly. Because the ordinary holder/background is already neutralized, the appropriate experiment is not a basic slab-mask comparison but a residual-cue ablation: progressively erode a very small rim-adjacent boundary or otherwise remove possible fringe/reflection cues, then rerun the frozen grade-by-grade comparison. Any such erosion must be conservative enough not to remove genuine rim information that belongs to the coin itself. This is a future instrument-validation study rather than part of the present confirmatory result.

A stronger future design is issue-blind. Date and mint information should be masked or otherwise prevented from entering the measurement system, with strike quality measured independently where possible. The analysis can then compare the same nominal grade across date/mint issues and construct issue-specific ladders such as MS-63 to MS-64 to MS-65 to MS-66. If a 1921 MS-65 and an 1881-S MS-65 occupy systematically different positions under an issue-blind ruler, that would be direct evidence that the market label is issue-relative in apparent visual quality. This remains a high-value future study, but it should not delay release of the current paper.

8.10 Relation to common Morgan-grading beliefs

The current evidence supports some common collector intuitions, rejects simple versions of others, and leaves several long-standing claims genuinely unresolved. The table below summarizes only what follows from this study.

Common claim	What this study shows	Assessment
PCGS grades Morgan dollars tougher than NGC across the board.	Service separation is strongly grade-dependent: largest near AU-55/AU-58, near parity in primary MS-62 through MS-66, with some issue cells favoring NGC.	Simple universal claim not supported.
PCGS is generally more consistent than NGC.	Overall within-cell dispersion is very similar. MS-66 shows a meaningful tighter-PCGS pattern, but no broad service-wide advantage appears.	Not supported as a general rule.
“Buy the coin, not the holder”: two coins with the same date, mint, service, and grade need not be equivalent in apparent quality.	Same-label populations retain nontrivial model-score dispersion; in mid-MS, random same-label pairs differ by >=1 modeled point roughly 13-27% of the time.	Supported in this narrow population sense; holder grade does not imply visual interchangeability.
Certain dates/grades are "guarded."	Large issue-specific candidates exist in both	Suggestive candidates only, not established.

Common claim	What this study shows	Assessment
	directions, but none survived the frozen BH-FDR threshold and all were heavily source-concentrated.	
The AU-58/MS-60 boundary behaves differently from the commercial MS core.	The confirmatory pattern and external cohort both point to unusual service separation around upper AU / low MS, followed by convergence through much of MS-62 to MS-66.	Supported as a population pattern; mechanism unknown.
Key dates are graded against a different quality standard.	Among the 207 well-populated matched cells, commercial MS grade centers vary only modestly by date/mint within each service. At MS-65, issue-mean SD is 0.16 points in both services. However, rare/key cells that fail n>=30 per service are not tested, and rarity itself was not modeled.	No broad issue-specific shift supported in well-populated commercial MS; true key-date behavior remains open.
Weakly struck dates receive a different grading allowance.	Strike was not independently measured. Only 1921-P MS-64/MS-65 met eligibility and their common shifts were small (-0.16, -0.24), but v9 may inherit issue conventions, so this does not test a strike allowance.	Open. Requires independent strike measurement and preferably an issue-blind ruler.
An MS-65 Morgan represents one abstract series-wide quality standard regardless of date/mint.	Under v9, MS-65 centers are quite stable across 38 eligible issues: range 0.74 PCGS / 0.71 NGC and issue-mean SD 0.16 in both. But the model learned from professional labels and is not issue-blind.	Current holder populations are close to issue-invariant at MS-65 under v9; abstract issue-blind invariance is not established.

9. Limitations

Limitation	Implication
Representativeness	The corpus is a large dealer-sourced observational sample, not a random census of every Morgan dollar certified by each service.
CV-verification selection	The primary cohort is selected by slab-certification verifiability. NGC passed the gate more often than PCGS. The external held-out cohort directly tests but does not eliminate every selection concern.
Frozen cohort scope	The v9 cohort predates a later archive-pipeline repair. Separately, a closeout audit found 91 modern physical coins that violated the preregistered classic-date scope; they were mechanically removed, changing headline estimates by at most 0.001.
Submission and market selection	Coins are not randomly assigned to services. Submission, crossover, crackout, reconsideration, and resubmission behavior can create population differences even if initial grading standards were identical.
Unknown grading date	The original grading date is generally unavailable. Holder generation was deliberately not used as grading date because reholdering breaks that inference.
Human graders and local eras	Smaller issue/grade cells may reflect particular graders or local periods rather than stable company-wide practice.
Training-label dependence	OOF prevents physical-coin memorization but does not make the model independent of the professional grading labels in the broader training corpus. A service-balanced or service-reweighted training robustness variant remains desirable if computationally feasible.
Holder and photography cues	Direct holder identifiers are removed before inference. Small residual boundary, reflection, or acquisition cues cannot be excluded; see Section 4.5 for the complete treatment.

Limitation	Implication
Proprietary instrument	The statistical pipeline is reproducible from frozen outputs, but the instrument itself cannot be independently reconstructed from the public paper without confidential model assets.
Issue heterogeneity	AU-55 controls were less uniform than AU-58 controls. A single overall service effect does not describe every date/mint cell.
MS-68	The primary MS-68 sample was too small for useful inference.
External metadata verification	The external cohort has no recorded CV mismatch but lacks positive CV agreement. It is fully nonoverlapping with v9 train/validation/test and is valuable as a sensitivity cohort, but its certification metadata are less independently verified than the primary cohort.
Cross-service re-certification linkage	The identity system can collapse repeated appearances of the same slab/certification but cannot exhaustively identify a physical coin cracked out or crossed into another service under a new certification number. Residual dependence across service populations is therefore possible.
Upstream pHash over-deduplication	A manual audit of 100 frozen Morgan pHash drops found 53 false merges, 17 correct duplicate removals, 1 ambiguous case, and 29 unresolved cases. The error removed legitimate listing records from one archive source rather than contaminating retained records with foreign labels; extrapolation suggests roughly 4,800 Morgan rows may have been over-dropped as coverage loss. No meaningful certification-service direction bias was detected.
Cross-side uncertainty robustness	The predeclared uncertainty analysis was not completed because the frozen plan did not operationally specify a unique subset rule and effect measure. No post hoc implementation was substituted.
Model-family dependence of replication checks	The old and new graders share development lineage and historical training-label environment, and the five external fold-model pairs are separate fits with heavily overlapping training data. These analyses test robustness to instrument revision and model fit, not replication by five statistically independent instruments.
Exploratory issue-cell source concentration	Every prespecified issue-outlier candidate was dominated by one confidential source, often >=90% of both service populations. Within-source direction agreed, but generalization beyond that inventory is limited.
Exploratory multiplicity	No issue-specific excess-PoS candidate survived Benjamini-Hochberg FDR at q<0.05. Candidate guarded-grade patterns are hypothesis-generating replication targets, not confirmed discoveries.
Internal-consistency estimand	Dispersion compares current same-issue holder populations under the common visual ruler. It does not directly measure repeatability of individual human graders or grading desks.
Absolute-dispersion interpretation	Within-label SD and pairwise score differences are properties of the current holder populations under the common instrument. They are not estimates of grading-service error or distance from an objective true grade.
Nonuniform Sheldon spacing	Raw Sheldon-point dispersion is directly comparable to adjacent grade steps in much of Mint State but not in circulated grades, where neighboring labels can be 5, 10, or 20 numerical points apart. Cross-grade comparisons should use grade-step normalization or another local scale.
Repeat-image diagnostic	Only 63 same-coin, same-side repeat-image units were available, all within the PCGS population. Their median absolute re-score difference was 0.26 Sheldon points (mean 0.44; 90th percentile 1.03), which is smaller than typical mid-Mint-State same-label coin-to-coin differences and argues against measurement wobble explaining the entire spread. The sample is too small and unrepresentative for formal variance decomposition, does not test

Limitation	Implication
	systematic bias, and should not be used to claim a precise percentage of “real” coin variation. A purpose-built repeatability study is being developed for separate future release.
Issue-specific standards and strike	Between-issue centers have now been measured under v9 and are relatively stable in well-populated commercial MS, but weak strike, luster norms, die state, and issue-specific allowances were not independently annotated. Rare/key cells failing the 30/30 threshold remain outside inference.
Issue-specific label inheritance by the instrument	Because the common grader learned from professional grading labels, it may inherit date/mint-specific grading conventions. A direct test of abstract versus issue-conditioned standards should use an additional issue-blind or date/mint-masked design.
Issue-relative grading standard / weak strike	The current v9 center-shift analysis cannot distinguish a truly generic standard from an issue-relative convention inherited by the model. Only an issue-blind ruler with independent strike measurement can directly test whether weak-struck or difficult issues receive allowances.

10. Reproducibility, Registration, and Data Availability

The confirmatory specification was frozen and registered before the new-grader service results were inspected. The canonical OOF file, identity map, leakage exclusions, original confirmatory script, primary output manifest, external output manifest, each corrective analysis, and the final classic-scope cohort were hashed. Nine implementation omissions were documented as they were discovered. Each corrective implemented an analysis requirement that predated result inspection or mechanically enforced the preregistered scope; no corrective introduced a new hypothesis, threshold, or discretionary exclusion. The paper and video should draw numbers from the final classic-scope corrective outputs wherever an earlier output was affected. The subsequent consistency/outlier exploration was governed by a separate specification frozen before exploratory result inspection (SHA-256 9e3c2f2f55398dace90787a4574a2da98614312df9deafbcb891dd82e6484820); its script and run manifest were also hashed before interpretation. The absolute same-label dispersion follow-up was governed by a second frozen specification (SHA-256 7a420ab655b1cea46ec73f1261de5a6b7b7485e1ec2a485a435cfdc74fbb61e4); its analysis script was frozen before execution (SHA-256 6d22998ad6a3d74059523fe1b227e528d70dc43b5afbab27b00624cc6d44299d), and its summary, repeat-image diagnostic, and run manifest were hashed before interpretation. The date/mint-average stability follow-up was governed by a third frozen exploratory specification (SHA-256 680ca792a702930dadc0d26524035986679c992abaf7a8ae8def121bb479f774), with analysis script frozen before execution (SHA-256 71003f40e586dcab023bb09e9cd1227e0a770bbbe94a9d59621ca69e9b6792f2) and output manifest SHA-256 fb8bb0763d1a78a58dbfc9ab9dc1e303b8a2240377bac306626b337f26e42e2f. The separate planned repeatability study is not part of the frozen confirmatory or exploratory claims reported here and will be released independently when complete.

Main-text tables report rounded values from the final classic-scope outputs. Full-precision estimates, confidence-interval endpoints, per-file hashes, and the tiny MS-65/MS-66 scope corrections are retained in the frozen CSV files and manifests listed below.

Registration. Lowry, T. “Confirmatory Analysis of Third-Party-Graded Morgan Dollar Populations.” Open-Ended Registration, OSF Registries, registered 21 August 2026 at 09:58:41 EDT [18]. The record was under embargo at the time this preprint was prepared; the title, timestamp, and frozen specification hash are reported here so the registration can be identified when public.

Data and code availability. Frozen specifications, analysis scripts, hash manifests, and de-identified coin-level output tables are available from the author for noncommercial research review, subject to confidentiality and commercial constraints. Raw dealer imagery, dealer identities, confidential source labels, model weights, and proprietary

preprocessing/training assets are not available. A public reproducibility package is planned after final confidentiality review.

Artifact	SHA-256
Registered v2 specification	25cf6b096a707756cf42f328c5df6ac923f7c2d2b3fe05dcea40e65128bfadf7
Canonical fp32 OOF file	3018f835d59830142fbb005e6cb81d21834caa21dcd31930d780e0c4b6b30bdb
Final identity map v2	33e12cdffb13c9e89211de2b27bb22709c1cab2782a4e7eb8a14a945e676ecd3
Physical-leakage exclusion list	640ebf90c8642ba9b259b33f865d6721971664b5912c6f9d112d1f21cd532abe
Original confirmatory script	e2d4ed5eea17d0a9fad53eff9530652d8d808ecaea324fd8d8cd7d7a4d9f97d4
Primary outputs manifest	f1aad8dca4f86df3f33a8d7978e5148d262efb51d154775f5f93862794c72b5a
External outputs manifest	29180d94a2970888eb62ee35ab336db9bcd97875ba1c99bf3574c1150e90d63c
External execution fix note	85ab730b979931d0899a4f2821fbf310c46ab2765b7e634dec3ab949f7e4981b
Revised external-execution script	7017d57a860ff799d154913a1585f3ba54d5415fee28f1bd322219199fd3fd34
Original/revised script diff audit	dfd66e9eba553fdeaa8744c775040b3caf42c973587603eef7ddb269f4437d7d
Probability-vector physical-coin corrective output	fb4e03a12be33a690add4572952b6a5fcb009f7463af1552b1d1207bb88ec1ff
Both-sides corrective manifest	7c21aed051f9777c0e6a68a22ed86e73ded27c16b02f4705dd4a3dfab67200f9
Classic-scope omission note #9	5f0c9629cb5949ac60d6b5c2a0dd903414ff20f716cd51cab95824a08e872842
Classic-only analysis-coin ID list	7d4cdf1e5746a54047fbc4b7b232c5c4054d47765f8d192e55836d78d32d0a6f
Classic-scope comparison table	037880598aa2327d6d0aa8a725a4a24c5c920a1f449abb78e1b66f1f7232857c
Classic-scope corrective manifest	f0d098d0e96f74b7dbb0f2e0c94c46d7ec5b3c8995d2263965e37c1bdde01eaf
Exploratory consistency/outlier specification	9e3c2f2f55398dace90787a4574a2da98614312df9deafbcb891dd82e6484820
Exploratory consistency/outlier script	a55d3b36a2d2a77c062403fdf4d9074228ad2f17f8bda9c128819032e717323d
Exploratory run manifest	b79358f2352ad2b66ba5d489e70108cec57b02632a6039315b91b3f57788a716
Exploratory candidate table	34a8a422aa5f64656f854112f4d53a95e7fe38a192640e86d4811d2c964ab821
Absolute-dispersion exploratory spec	7a420ab655b1cea46ec73f1261de5a6b7b7485e1ec2a485a435cfdc74fbb61e4
Absolute-dispersion analysis script	6d22998ad6a3d74059523fe1b227e528d70dc43b5afbab27b00624cc6d44299d
Absolute-dispersion summary output	690f4b667de27b7c2c98559597fdaf9bb4525b2061998b4135eef1da79d13086
Repeat-image diagnostic output	b0674622b4fa00330065d6a058279ade636332a44710fc3d575ac084b

Artifact	SHA-256
	91f18ce
Absolute-dispersion run manifest	e6a969ed5c18de03b086b31b272e3be6a64106737ed5b6af59655c8b7e2fad49
Issue-center invariance exploratory specification	680ca792a702930dadc0d26524035986679c992abaf7a8ae8def121bb479f774
Issue-center invariance analysis script	71003f40e586dcab023bb09e9cd1227e0a770bbbe94a9d59621ca69e9b6792f2
Issue-center invariance output manifest	fb8bb0763d1a78a58dbfc9ab9dc1e303b8a2240377bac306626b337f26e42e2f

The full output manifests retain per-file hashes. Raw dealer imagery, dealer identities, model weights, and proprietary implementation are not intended for public release. The intended public reproducibility unit is the frozen coin-level output package, analysis code, inclusion and aggregation rules, grade mappings, and generated summary tables, subject to confidentiality and commercial constraints.

11. Competing Interests and Author Disclosure

The author has a financial interest in CoinAiLyzer and in the proprietary grading technology used as the measurement instrument in this study. CoinAiLyzer is not the subject of the paper, but its grading engine generated the model outputs analyzed here. The relationship is disclosed so readers can evaluate the work with full knowledge of the commercial interest.

12. Conclusion

A preregistered confirmatory analysis using an improved computer-vision grader met the registered Strong Replication definition in the primary cohort and demonstrated a highly non-uniform PCGS-NGC population difference among classic Morgan dollars. PCGS AU-55 and AU-58 Morgans scored materially higher than NGC coins with the same stated grades, while the primary MS-62 through MS-66 populations met a practical-parity criterion. A fully held-out external cohort reproduced upper-AU separation and MS-core convergence but peaked at MS-60 rather than AU-58, showing that the exact boundary peak is cohort-dependent. The broader upper-AU/low-MS boundary pattern survived same-coin old/new measurement on essentially identical images, both faces, the both-sides-present restriction, cert-anchored identity restriction, source deletion, and external scoring. Separately frozen exploratory analyses found no uniform service-wide consistency advantage and no issue-specific service outlier surviving BH false-discovery control. Same-year, same-mint, same-grade, same-service populations nevertheless remain meaningfully heterogeneous: median within-label SD was 0.63 Sheldon points for PCGS and 0.65 for NGC, and median-cell probabilities of a >=1-point random-pair difference were 24% and 26%. Commercial Mint State grade averages were comparatively stable across well-populated date/mint issues; at MS-65, 38 eligible issues spanned only 0.74 points in PCGS and 0.71 in NGC, with an issue-mean SD of 0.16 in both. Thus the data support “buy the coin, not the holder” more strongly than they support the idea that every well-populated Morgan issue requires a substantially different Mint State quality standard. This does not settle key-date grading or weak-strike allowances. Strike was not independently measured, rare issue cells often fail the eligibility threshold, and the model may inherit issue-specific professional conventions. A future issue-blind benchmark with independent strike measurement is the appropriate test of whether generic published standards are truly invariant across Morgan issues.

References

[1] Professional Coin Grading Service. PCGS Grading Standards. PCGS. https://www.pcgs.com/resources/pdf/PCGSGradingStandards.pdf (accessed 22 August 2026).

[2] Numismatic Guaranty Company. NGC Coin Grading Scale. NGC. https://www.ngccoin.com/coin-grading/grading-scale/ (accessed 22 August 2026).

[3] PCGS CoinFacts. 1921 $1 Morgan (Regular Strike), Morgan Dollar. Commentary on characteristic weak strike and cross-date appearance differences. https://www.pcgs.com/coinfacts/coin/1921-1-morgan/7296 (accessed 22 August 2026).

[4] PCGS CoinFacts. 1921-S $1 (Regular Strike), Morgan Dollar. Commentary on weak strike and comparatively loose grading interpretations. https://www.pcgs.com/coinfacts/coin/1921-s-1/7300 (accessed 22 August 2026).

[5] Sheldon, W.H. Early American Cents, 1793-1814. New York: Harper & Brothers; 1949.

[6] Zambanini, S.; Kampel, M.; Schlapke, M. On the Use of Computer Vision for Numismatic Research. In: Eurographics Workshop on Cultural Heritage; 2008:17-24.

[7] Kavelar, A.; Zambanini, S.; Kampel, M.; Vondrovec, K.; Siegl, K. The ILAC-Project: Supporting Ancient Coin Classification by Means of Image Analysis. Int Arch Photogramm Remote Sens Spatial Inf Sci. 2013;XL-5/W2:373-378. doi:10.5194/isprsarchives-XL-5-W2-373-2013.

[8] Guo, Z.; Arandjelovic, O.; Reid, D.; Lei, Y.; Buttner, J. A Siamese Transformer Network for Zero-Shot Ancient Coin Classification. J Imaging. 2023;9(6):107. doi:10.3390/jimaging9060107.

[9] Chen, J. Development of an Automated Coin Grading System: Integrating Image Preprocessing, Feature Extraction, and ML Modeling. Master's thesis, Virginia Tech; 2024. http://hdl.handle.net/10919/123863.

[10] Getahun, M.N.; Anikin, D.; Shmatok, A.; Korchagin, M.; Zaytsev, A.; Somov, A. Dual-Image Vision-Language Framework with Expert Knowledge for Automated Fine-Grained Mint-State Coin Grading. Expert Systems with Applications. 2026;331:133156. doi:10.1016/j.eswa.2026.133156.

[11] Getahun, M.N.; Somov, A. Enabling Automatic Coin Grading Through Late-Fusion of Vision-Language and Hierarchical Transformers. IEEE Transactions on Instrumentation and Measurement. 2026;75. doi:10.1109/TIM.2026.3699676.

[12] Mann, H.B.; Whitney, D.R. On a Test of Whether One of Two Random Variables Is Stochastically Larger than the Other. Ann Math Stat. 1947;18(1):50-60. doi:10.1214/aoms/1177730491.

[13] Vargha, A.; Delaney, H.D. A Critique and Improvement of the CL Common Language Effect Size Statistics of McGraw and Wong. J Educ Behav Stat. 2000;25(2):101-132. doi:10.3102/10769986025002101.

[14] Cohen, J. Statistical Power Analysis for the Behavioral Sciences. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.

[15] Efron, B.; Tibshirani, R.J. An Introduction to the Bootstrap. New York: Chapman & Hall; 1993.

[16] Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc B. 1995;57(1):289-300.

[17] Nosek, B.A.; Ebersole, C.R.; DeHaven, A.C.; Mellor, D.T. The Preregistration Revolution. Proc Natl Acad Sci USA. 2018;115(11):2600-2606. doi:10.1073/pnas.1708274114.

[18] Lowry, T. Confirmatory Analysis of Third-Party-Graded Morgan Dollar Populations. Open-Ended Registration, OSF Registries; registered 21 August 2026, 09:58:41 EDT. Registration under embargo at the time of manuscript preparation.

Appendix A. Confirmatory Registration and Pre-Result Amendments

The v2 specification was frozen at 08:44 EDT and registered on OSF at 09:58:41 EDT on 21 August 2026. A pre-result amendment documented data-integrity decisions made after registration but before service-score inspection: acceptance of the frozen v9 cohort, use of service-consistent certification identity with coin_key fallback, exclusion of 148 leaked physical coins, retention of canonical fp32 OOF, treatment of the CV-verified OOF as primary, creation of the external no-agree/no-mismatch cohort, descriptive-only treatment of MS-68, and addition of a cert-anchored identity sensitivity.

The frozen methods record also documents two diagnostics deliberately not used in the confirmatory analysis. Service predictability from visible coin crops was not treated as a confound test because it cannot distinguish holder

or photography cues from genuine service-population differences. Holder generation was not used as a proxy for original grading date because reholdering can place an older grading decision into a newer holder. These were pre-result exclusions, not post-result choices.

The analysis pipeline then revealed nine implementation omissions or scope mismatches. Each was documented in a separate note before its corrective computation, followed by a narrowly scoped script hashed before execution. The later corrections included physical-coin probability-vector aggregation, the missing both-sides-present robustness analysis, and the final enforcement of the preregistered classic-date scope after 91 modern physical coins were found in the clean cohort. The classic-scope correction affected only one NGC MS-65 coin and one PCGS MS-66 coin in the headline AU-50 through MS-67 range and changed no qualitative result.

No.	Implementation / audit issue	**Resolution**
1	Direct grade-shape table emitted raw-gap arithmetic rather than registered d and PoS contrasts with bootstrap CIs.	Corrective direct-contrast script.
2	Year/mint bucket d and LOSO d/PoS were omitted.	Corrective control-effect script.
3	Controlled PoS used an unweighted mean of stratum PoS rather than the registered pair-count estimator.	Corrective pair-weighted PoS script.
4	Primary table omitted bootstrap CIs for Cohen's d.	Corrective d-CI script.
5	Old/new matched analysis omitted service-gap, d, PoS, paired change intervals, correlations, and slopes.	Corrective old/new script after image-membership audit.
6	Side-specific analysis used image rows rather than one side-level mean per physical coin.	Corrective coin-level side script.
Exec.	External scoring hit a Windows DataLoader execution failure after the primary outputs had already been generated, hashed, and frozen.	Execution-only fix moved the dataset class to module scope / used safe worker settings. The original primary outputs remained untouched; a reconstruction and code-diff audit verified the primary statistical functions were unchanged.
7	Probability-vector diagnostic used image rows rather than the registered physical-coin aggregation.	Documented omission #7; corrective averaged vectors within side and then equally across sides at the physical-coin level. AU-55 and AU-58 diagnostics were then recomputed.
8	No frozen both-sides-present robustness output existed despite the explicit registered requirement.	Documented omission #8; corrective restricted to physical coins with both sides, retained the registered side aggregation, and reproduced the headline localization.
9	Ninety-one modern Morgan physical coins dated 2021-2025 remained in the clean cohort despite the preregistered classic-date scope.	Documented omission #9; mechanically restricted year to 1878-1904 or 1921, audited every registered output for impact, and recomputed only affected outputs. Headline changes were at most 0.001.

The existence of these corrections should be reported plainly rather than hidden. The important safeguards are that the registered requirements predated result inspection, each omission was identified from comparison with the frozen specification, each corrective was restricted to the missing requirement, and each corrective note and script was hashed before execution.

Appendix B. Exploratory Consistency and Issue-Outlier Follow-Up

Release boundary: the current confirmatory and three frozen exploratory follow-ups are complete enough to report. The between-date/mint center-shift analysis described earlier as high-value follow-up has now been completed under a separately frozen specification. The issue-blind standard-invariance benchmark, independent weak-strike measurement, secondary-service work, and same-coin crossover studies remain future research and are not prerequisites for release. This boundary is intentional to avoid extending the present study indefinitely through post hoc questions.

The confirmatory study was closed before these analyses began. The exploratory specification was frozen before any consistency or issue-outlier result was inspected (SHA-256 9e3c2f2f55398dace90787a4574a2da98614312df9deafbcb891dd82e6484820). All eligible cells required at least 30 unique physical coins per service. No issue cell survived the prespecified Benjamini-Hochberg q<0.05 threshold for excess PoS. The table below therefore lists the complete prespecified candidate set selected by the frozen rule: any q<0.05 plus the ten largest positive and ten largest negative excess-PoS cells, deduplicated.

Issue	Grade	n P/N	d	PoS	Excess PoS	BH q	Sides	Old/new
1882-CC	MS-64	51/34	+0.684	0.694	+0.190	0.116	same	disagree
1878-P	MS-61	47/36	+0.597	0.703	+0.183	0.116	same	agree
1887-O	MS-62	42/40	-0.609	0.330	-0.171	0.237	same	agree
1892-P	AU-55	36/41	-0.341	0.414	-0.159	0.406	same	agree
1894-O	AU-58	54/50	-0.165	0.450	-0.146	0.339	same	disagree
1902-S	MS-63	45/32	-0.542	0.356	-0.145	0.419	same	agree
1893-CC	EF-40	36/30	-0.681	0.311	-0.139	0.476	same	agree
1880-CC	MS-65	160/62	+0.541	0.644	+0.138	0.116	same	agree
1880-CC	MS-64	53/31	+0.498	0.642	+0.138	0.419	same	agree
1896-P	MS-66	238/96	-0.457	0.360	-0.137	0.083	same	agree
1904-S	VF-20	44/32	+0.235	0.617	+0.128	0.476	opposite	disagree
1890-CC	MS-64	137/36	+0.497	0.633	+0.128	0.414	same	agree
1898-O	MS-63	47/40	+0.430	0.617	+0.118	0.476	same	agree
1901-O	MS-66	83/78	+0.283	0.601	+0.117	0.313	same	agree
1888-S	MS-64	89/34	+0.359	0.617	+0.112	0.476	same	disagree
1886-O	AU-50	43/35	-0.286	0.414	-0.110	0.488	same	disagree
1882-S	MS-67	56/108	-0.208	0.443	-0.109	0.419	same	agree
1921-P	MS-65	47/43	+0.354	0.614	+0.106	0.476	same	agree
1885-P	MS-62	38/40	-0.429	0.399	-0.101	0.497	same	agree
1886-O	AU-58	50/47	+0.086	0.494	-0.098	0.488	opposite	agree

Interpretation: 18 of 20 candidates had the same direction on obverse and reverse, and 15 of 20 had the same direction under v8 and v9. Every candidate was highly concentrated in a single confidential source. These are issue-specific service-population outliers suitable for independent replication, not evidence of deliberate grade guarding.

Appendix C. Absolute Same-Label Dispersion Follow-Up

This exploratory descriptive analysis was frozen before result inspection under specification SHA-256 7a420ab655b1cea46ec73f1261de5a6b7b7485e1ec2a485a435cfdc74fbb61e4 and executed with a pre-run frozen script SHA-256 6d22998ad6a3d74059523fe1b227e528d70dc43b5afbab27b00624cc6d44299d. It reused the 207

issue cells from the consistency analysis. Equal-cell median SD was 0.63 Sheldon points for PCGS and 0.65 for NGC. Median-cell probabilities of a random same-label pair differing by at least 1.0 point were 0.24 and 0.26, and by at least 2.0 points were 0.033 and 0.034. These are common-instrument dispersion measures, not grading-error estimates.

The limited repeat-image diagnostic contained 63 same-coin, same-side repeat-image units, all PCGS. Median absolute repeat-image difference was 0.26 Sheldon points, mean 0.44, and 90th percentile 1.03. Mid-Mint-State same-label pair differences were materially larger than the repeat-image median. This argues against ordinary re-imaging/model wobble explaining the full same-label spread, but the repeat set is too small, one-service, and unrepresentative for formal noise subtraction or a precise estimate of the fraction attributable to coin versus measurement. A purpose-built repeatability study is being developed for separate future release; high-confidence linked repeats retained from deduplication may provide a complementary natural-repeat sample.

Interpretive caution: raw Sheldon points are not equally spaced grade steps below Mint State. The apparently very large raw SDs in VF, EF, and low AU partly reflect the numerical construction of the Sheldon labels. A future grade-step-normalized analysis is recommended before making cross-grade statements about which parts of the scale are intrinsically most variable.

Appendix D. Date/Mint Average-Stability Follow-Up

This final exploratory descriptive analysis was frozen before result inspection under specification SHA-256 680ca792a702930dadc0d26524035986679c992abaf7a8ae8def121bb479f774 and executed with pre-run frozen script SHA-256 71003f40e586dcab023bb09e9cd1227e0a770bbbe94a9d59621ca69e9b6792f2. Output manifest SHA-256: fb8bb0763d1a78a58dbfc9ab9dc1e303b8a2240377bac306626b337f26e42e2f. It reused the same 207 eligible year x mint x stated-grade cells and asked whether the center of a fixed grade shifts across date/mint within each service.

Commercial Mint State centers were relatively stable. PCGS/NGC ranges of eligible issue means were: MS-63 1.21/0.91 points, MS-64 0.78/0.75, MS-65 0.74/0.71, and MS-66 0.53/0.56. Corresponding SDs of issue means were MS-63 0.26/0.22, MS-64 0.19/0.18, MS-65 0.16/0.16, and MS-66 0.18/0.18. Across all 207 within-grade-centered cells, PCGS and NGC issue shifts had Pearson r=0.657 and 82.1% sign agreement.

Only 1921-P MS-64 and 1921-P MS-65 cleared the 30/30 eligibility rule. Their common issue shifts were -0.16 and -0.24 points. These small v9 shifts do not support a large visible downward center shift for the eligible 1921-P populations, but strike was not measured and v9 may inherit the same issue-relative conventions used in professional labels. The issue-blind design described in the Discussion remains the appropriate future test.