VERSION 0.2 / REVISED JULY 2026

Evidence-Based Marketing Playbook

A public-data replication audit of Ehrenberg-Bass regularities

Five claims from version 0.1 are withdrawn after a code audit.

Kyo Harada
Original analysis: September 2025
Revised public edition: July 2026

Revision notice - July 2026

This is version 0.2. The original version 0.1 (September 2025) is archived and remains available.

Re-auditing the original analysis code and logs against the published text showed that several statements in v0.1 exceeded what the code measured. Five claims are withdrawn:

1. **Top-decile Double Jeopardy deviation.** No top-decile or middle-quantile split exists in the script. The reported r=0.627 was computed across all 27 retained labels.
2. **R^2=0.472 as response to an additional contact.** The regression contains no contact, campaign, or treatment variable. It measures adjacent-quarter frequency persistence only.
3. **CEP coverage rising from 38% to 52%.** No before/after comparison exists in the archived output, and the associated <5% Δaccuracy statement is removed with it.
4. **Language-bias detection across 27 languages.** The logged run retained one language and used ASINs as brand identifiers.
5. **Poor Dirichlet model fit.** The evaluation path predicted the sample mean for every user, so R^2 approx. -7x10-6 evaluates a constant-mean predictor, not the fitted NBD distribution.

Implementation defects were also found in the bootstrap, negative-control, and NBD evaluation paths. The remaining numerical outputs are retained as an audit trail. They are not presented as a validated replication of the Ehrenberg-Bass laws.

What v0.1 got right is retained: the weighted Duplication of Purchase statistic did not cross its pre-set gate, and the simplified unweighted statistic was not substituted to manufacture a pass.

Measured states

DoP weighted MAD 0.015863 - FAIL (>0.015) / DJ Pearson r 0.627 - FAIL (<0.80) / Q4 lagged-frequency R^2 0.472 - descriptive only / CEP pipeline - not validated

Executive Summary

The 2025 project should be read as a replication audit, not as a validated marketing playbook. Its strongest property is that the archived decision gates preserve unfavorable results. Its principal weakness is that several public claims exceeded what the code measured.

Analysis	Archived result	Gate or check	Defensible conclusion
Duplication of Purchase, dunnhumby	weighted MAD 0.015863	<=0.015	Near-miss; the run failed
Duplication of Purchase, Instacart	weighted MAD 0.021854	<=0.015	Failed
Double Jeopardy, UCI	Pearson r=0.627	>=0.80	Failed; stationarity also failed
Buyer-frequency persistence, UCI Q4	R^2=0.472	no confirmatory gate	Descriptive association only
CEP lexical pipeline, Amazon	r=-0.280	no valid language-bias test	Not interpretable because the parser and configuration schemas disagree
NBD diagnostic, UCI	R^2 approx. -7x10-6	no valid goodness-of-fit test	Not interpretable as NBD fit because predictions were set to the sample mean

Three claims in the previous edition are withdrawn.

1. **There was no top-decile Double Jeopardy analysis.** The value r=0.627 was computed across all 27 retained labels.
2. **R^2=0.472 does not measure response to an additional contact.** The model regressed transaction counts across adjacent quarters and included no contact, campaign, or treatment variable.
3. **The archive does not show CEP coverage improving from 38% to 52%.** The logged CEP run retained one language, used ASINs as brand identifiers, and produced no before/after comparison.

The negative results still matter. The weighted DoP calculation did not cross its gate, and the simplified unweighted calculation cannot be substituted to manufacture a pass. That reporting discipline should be retained when the analysis is rebuilt.

The current asset is not decision-ready or publication-ready. Publication requires corrected estimators, versioned analysis code, redistributable or checksum-verified inputs, and an independent rerun from a clean environment.

What We Measured

The project combined four datasets that differ in sampling frame, market definition, and unit of observation. Their outputs cannot be pooled as if they described one market or one period.

Dataset	Role in the archive	Logged analysis sample
dunnhumby Complete Journey	Duplication of Purchase	1,735 households, 46 retained labels
Instacart Online Grocery Shopping 2017	Duplication of Purchase comparison	1,617 users, 129 retained labels
UCI Online Retail II	Double Jeopardy, buyer-frequency persistence, NBD diagnostic	1,264 users and 27 labels for DJ; 1,666 users for the quarterly analysis; 1,648 users for NBD
Amazon Review Data (2018)	lexical CEP prototype	1,000,000 review rows processed; 58 ASINs in the retained output

Research questions

1. Duplication of Purchase

The intended question was whether cross-brand buyer duplication remains close to a common level in a repertoire market. The project summarized off-diagonal duplication values with a weighted mean absolute deviation and compared the result with an internal gate of 0.015.

2. Double Jeopardy

The UCI run correlated each retained label's buyer penetration with its average transaction frequency among buyers. The internal gate required Pearson r>=0.80 and a lower confidence bound of at least 0.70.

3. Purchase-frequency persistence

Users were assigned to purchase-volume quartiles within each calendar quarter. Within each quartile, the script regressed one quarter's transaction count on the adjacent observed quarter's transaction count. This measures within-user frequency persistence under the script's grouping rule. It does not measure response to advertising, reminders, or incremental contact.

4. CEP lexical coverage

The intended construct was the prevalence of Category Entry Point language in product reviews. The implemented run searched a multilingual lexicon, aggregated hits by ASIN and detected language, and correlated average lexical coverage with review count. A configuration-schema mismatch prevents the output from measuring the intended CEP dimensions or multilingual coverage.

What was not measured

The project did not observe a 2025 Q3 client market, did not run a marketing intervention, did not estimate incremental sales, did not compare a 38% baseline with a 52% post-treatment value, and did not test Double Jeopardy separately in middle and top buyer quantiles.

Finding 1: Duplication of Purchase Did Not Cross the Gate

The closest archived run used a 26-week dunnhumby beauty slice after retaining users at or above the 0.90 purchase-count quantile, requiring at least two labels per user, and requiring at least 20 buyers per label. The final matrix contained 1,735 users and 46 labels.

Metric	Result	Gate	Decision
Weighted MAD	0.015863	<=0.015	FAIL
Gap to gate	+0.000863	-	Near-miss
Unweighted MAD	0.015143	not the confirmatory metric	No decision
Logged interval	[0.014792, 0.016928]	-	Descriptive only
Median labels per user	2.0	>=2.0	Met
Negative-control MAD	0.015629	<=0.05	Met under the project rule

The Instacart comparison was less favorable: weighted MAD=0.021854 on 1,617 users and 129 labels. Its unweighted MAD was 0.011934. The lower unweighted value cannot replace the pre-specified weighted statistic after the result is known.

Interpretation

The narrow conclusion is that no archived specification-labelled run crossed all gates. Calling 0.015863 a near-miss is accurate; calling it a pass is not.

The distance from the gate does not show that the Duplication of Purchase law is almost verified. Fourteen filter combinations were examined, and the reported run was selected as the closest result. The threshold is also an internal project rule, not a universal rejection boundary supplied by the underlying theory. A confirmatory replication must fix the category definition, observation window, inclusion rules, duplication estimator, and decision gate before examining the new data.

Implementation qualification

The current script copies one directional conditional duplication rate into both halves of a symmetric matrix. It also calculates weights after reducing the data to one user-label row, so the values called "purchase-count weights" are effectively retained buyer counts. The logged interval uses percentile resampling of matrix cells, not a bias-corrected and accelerated bootstrap of buyers. These defects prevent inferential use of the interval and require the point estimate to be treated as an archived pipeline output rather than a validated estimator.

Finding 2: The Double Jeopardy Run Failed Its Correlation and Stationarity Checks

The UCI run retained 27 labels and 1,264 users in a 26-week slice. Across those 27 labels, buyer penetration and average transaction frequency among buyers had Pearson r=0.6269 and Spearman r=0.5624.

Metric	Result	Gate	Decision
Pearson correlation	0.627	>=0.80	FAIL
Spearman correlation	0.562	descriptive	-
Pearson p-value	0.00047	not a fit gate	-
Logged stationarity	false	required for stable interpretation	FAIL
Maximum logged drift	0.375	<=0.10 under the script rule	FAIL

The result does not support the project's strong Double Jeopardy criterion. It also does not establish a failure of the empirical law. The UCI source is a UK giftware retailer, while the "bodycare" slice and label field were produced by heuristic category and brand transformations. The analysis therefore mixes a theoretical question with uncertain construct validity.

Correction to the previous edition

The script contains no middle-quantile or top-decile Double Jeopardy split. The value r=0.627 is the correlation across all retained labels. Claims that the relationship held in the middle but weakened among the top 10% are unsupported by this run.

Interval qualification

The logged interval [0.275, 0.462] does not contain the point estimate 0.627. Code inspection explains the mismatch: the bootstrap path computes penetration with a different denominator than the point-estimate path, and sampling users with replacement is reduced to set membership before recomputation. The interval is not used in this report.

Finding 3: High-Volume Buyers Showed Stronger Adjacent-Quarter Frequency Association

The quarterly UCI analysis assigned users to purchase-volume quartiles within each quarter, then regressed transaction frequency in one observed quarter on the adjacent observed quarter within each quartile. The predictor was standardized before estimation.

Purchase-volume quartile	Observations	Users	Standardized slope	R^2
Q1	126	107	-0.002	0.00001
Q2	88	76	0.321	0.196
Q3	77	67	0.765	0.204
Q4	190	120	3.341	0.472

Q4 had the strongest within-sample association. The result says that adjacent-quarter transaction counts were more predictable among the highest purchase-volume observations under this grouping rule.

What the model does not identify

There is no marketing-contact variable, treatment assignment, campaign exposure, stock measure, or price control in the regression. R^2=0.472 therefore does not mean that one additional contact causes repeat purchase, nor that the model explains 47% of incremental sales.

The script names the shifted value `freq_t1`, but the shift points to the previous recorded quarter. Quartile membership is calculated from purchase volume in each quarter rather than from a fixed pre-period. These choices allow outcome-related grouping and make the Q1-Q4 comparison partly mechanical. A causal or predictive follow-up must define a fixed baseline cohort, preserve time direction, hold out later periods, and introduce the exposure whose effect is being estimated.

Finding 4: The CEP Run Is a Failed Measurement Prototype

The archived Amazon run processed 1,000,000 review rows and wrote 256 aggregate rows covering 58 ASINs. It reported Pearson r=-0.280 between total review count per ASIN and mean lexical hit rate.

That coefficient is not evidence of language bias. The audit log shows one retained language (`en`) and 27 excluded language codes. The run therefore did not compare 27 languages.

Why the intended construct was not recovered

Two schema mismatches change the meaning of the output.

1. The CEP configuration is organized as `language -> dimension -> terms`, while the parser expects `dimension -> language -> terms`. The output consequently records `en` as the sole CEP category rather than quality, value, innovation, sustainability, and convenience.
2. The parser passes ASIN values into a nested brand-normalization dictionary that expects brand-name keys. The retained "brands" are therefore product identifiers.

The variable named penetration is total review count, not buyer penetration. The value r=-0.280 is a correlation between ASIN review volume and the malformed lexical-coverage output.

Withdrawn claim

No logged result or archived output contains a before/after test showing bottom-five CEP coverage rising from 38% to 52%. The previous edition converted two percentages into an intervention result without an observed intervention. That claim, and the associated `<5% Δaccuracy` statement, are removed.

NBD diagnostic

The archive also reports NBD parameters and R^2 approx. -7x10-6. The fitting routine estimates NBD parameters, but the evaluation path predicts the same sample mean for every user. The resulting R^2 evaluates a constant-mean predictor, not the fitted NBD distribution and not a full NBD-Dirichlet model. It cannot be used to claim that the Dirichlet model fits poorly.

Methods

This section describes what the archived scripts did. Labels such as "specification-compliant" and "BCa" are not adopted where the implementation does not meet those definitions.

Data preparation

- **dunnhumby:** household transactions were joined to product metadata, mapped to a beauty category, and filtered to the most recent 26 weeks in the processed file.
- **Instacart:** orders and products were joined, product names were mapped to shampoo labels, and the most recent 26 weeks were retained.
- **UCI Online Retail II:** invoice rows were transformed into user, date, quantity, category, and derived-label fields. The source describes a giftware retailer; the beauty/bodycare mapping is project-specific.
- **Amazon Review Data (2018):** up to 1,000,000 review rows were converted to TSV and processed in 100,000-row chunks.

Counts reported in this document are post-filter counts from the audit logs, not source-dataset totals.

Duplication of Purchase

The chosen dunnhumby run applied these conditions:

- category regex: `beauty`
- 26-week window
- user purchase-count quantile: >=0.90
- at least two retained labels per user
- at least 20 buyers per label
- random seed: 42

The script reduced the data to one row per user-label pair and formed a binary user-label matrix. For each off-diagonal pair it calculated a conditional overlap rate. It then copied the upper-triangle value into the lower triangle. The summary statistic was the weighted mean absolute deviation of upper-triangle values from their weighted mean, using the outer product of retained label counts as weights.

The script ran 5,000 percentile resamples of off-diagonal matrix cells, a within-user week permutation, and a random label-assignment negative control. The output calls the first procedure BCa. It does not calculate bias correction or acceleration and does not resample the primary sampling unit.

Double Jeopardy

For each retained UCI label, penetration was the number of unique buyers divided by total unique users. Average frequency was row count divided by unique buyers. Pearson and Spearman correlations were computed across labels.

The stationarity routine aggregated weekly penetration and frequency, measured first-to-last relative drift, and combined it with Kendall trend tests. The bootstrap routine attempted 5,000 user resamples, but replacement multiplicity was lost through membership filtering and its penetration denominator differed from the point-estimate denominator.

Buyer-frequency persistence

Transaction rows were aggregated to user-quarter observations. Purchase-volume quartiles were recomputed within each quarter. Within each quartile, the script standardized the current transaction count and regressed the shifted adjacent-quarter count on it using ordinary least squares. Reported confidence intervals are t-based slope intervals from that regression.

CEP lexical prototype

The parser detected review language, required at least 20 rows per ASIN-language cell, searched configured substrings, computed Wilson intervals for hit proportions, and correlated total review count with mean lexical coverage. The parser/configuration mismatch described in Finding 4 invalidates the intended dimension and multilingual interpretation.

Archived evidence

Run	Input identification	Timestamp
dunnhumby DoP	SHA prefix `e2470f224f7f0609`	2025-09-27 14:12:31
Instacart DoP	SHA prefix `fa486f16bde1d909`	2025-09-27 01:01:06
Amazon CEP	SHA prefix `faa6eadcba54534f`	2025-09-27 01:22:17
UCI DJ	recorded only as `loaded`; commit `unknown`	2025-09-23 00:55:12
UCI buyer frequency	recorded only as `loaded`; commit `unknown`	2025-09-23 13:28:06
UCI NBD diagnostic	recorded only as `loaded`; commit `unknown`	2025-09-23 13:28:45

The absence of hashes and commit identifiers for the UCI runs prevents exact provenance reconstruction.

Limits and Claim Boundaries

1. The datasets do not form one market

The analyses use grocery orders, household retail transactions, giftware invoices, and product reviews. Market boundaries, buyer identifiers, observation windows, and label construction differ. Agreement or disagreement across these outputs cannot be attributed to one common consumer process without a harmonized design.

2. Category and brand validity are unresolved

The UCI source is a UK giftware retailer. Its bodycare category and brand-like labels were created by project heuristics; retained labels include product-description tokens rather than verified beauty brands. The Amazon pipeline retained ASINs rather than normalized brands. These transformations weaken every claim that depends on a stable brand or category unit.

3. The DoP estimator requires reimplementation

- Directional conditional duplication was forced into a symmetric matrix by copying one direction.
- Purchase rows were deduplicated before the values used as weights were counted.
- The interval is a percentile interval over pairwise cells, not a BCa interval over independent buyers or households.
- The weekly shuffle check is applied after each user-label pair has been reduced to one row, making repeated weeks structurally unavailable.
- The best run was selected after multiple filter combinations were tried.

These conditions prevent the near-miss from functioning as a confirmatory statistical test.

4. The Double Jeopardy interval is invalid

The point estimate and bootstrap use different penetration denominators. Bootstrap draws with replacement are collapsed to membership, so repeated users receive no additional weight. The logged interval excludes the point estimate and is not interpretable. The stationarity check also failed.

5. The Q4 regression is non-causal

Purchase-volume quartiles are defined from contemporaneous outcomes, and the model contains no intervention. Higher Q4 R^2 can arise from the larger variance and persistence of high-volume buyers. No budget allocation or contact-frequency recommendation follows from this regression.

6. The CEP and NBD diagnostics do not test their named constructs

The CEP parser and lexicon schemas disagree, leaving one malformed category and one retained language. The NBD evaluation compares observations with a constant mean instead of predictions from the fitted distribution. Neither result can confirm or reject the corresponding theory.

7. Reproduction has not been demonstrated

The raw data are not bundled with the public report. Instacart and dunnhumby require account- or terms-mediated acquisition. The local analysis folder is not tied to a recorded Git commit, three logs record the input only as loaded, and the advertised one-hour end-to-end guarantee was never evidenced by a clean-room run.

8. There was no external review

No peer review, preregistration, or independent replication occurred. Thresholds were project-defined and should not be treated as universal theory tests.

Claim boundary

The archive supports only this claim: a 2025 exploratory pipeline recorded a near-miss DoP summary and several negative or uninterpretable diagnostics, while preserving enough logs to identify what must be rebuilt. It does not support causal marketing recommendations, validation of the Ehrenberg-Bass laws, or a published failure-to-replicate conclusion.

Reproduction and Publication Checklist

The current archive is not reproducible from the public report alone. This checklist defines the work required before the study can be described as a replication.

Freeze the research design

- [ ] Define each market and category without post-result filtering.
- [ ] Define the brand unit and publish the normalization table.
- [ ] Fix the observation window and buyer-inclusion rule before analysis.
- [ ] Justify each decision gate from prior literature or label it explicitly as exploratory.
- [ ] Separate exploratory parameter search from a held-out confirmatory run.

Repair the estimators

- [ ] Compute both directional duplication rates or use a symmetric overlap statistic whose denominator is defined in advance.
- [ ] Resample buyers or households, preserving replacement multiplicity.
- [ ] Implement an actual BCa interval, including bias correction and acceleration, or rename the interval.
- [ ] Run temporal controls on non-deduplicated event data and specify the null being tested.
- [ ] Use identical definitions in the Double Jeopardy point estimate and bootstrap.
- [ ] Define buyer quartiles from a fixed baseline period and evaluate later periods out of sample.
- [ ] Align the CEP parser with the lexicon schema and join ASINs to verified brand metadata.
- [ ] Evaluate NBD probabilities or quantiles generated from the fitted model rather than a constant mean.

Freeze provenance

- [ ] Place the analysis code in version control and record a commit SHA in every run.
- [ ] Record full SHA-256 hashes for every raw and processed input.
- [ ] Publish dataset acquisition dates, licenses, and transformation scripts.
- [ ] Lock the Python version and dependencies.
- [ ] Store result tables and figures outside an opaque archive with a manifest of hashes.

Validate from a clean environment

- [ ] Acquire each dataset from its documented source.
- [ ] Rebuild processed inputs without reusing local caches.
- [ ] Run unit tests on a small hand-checkable buyer-brand matrix.
- [ ] Recompute all primary statistics from one documented command.
- [ ] Confirm that a second operator obtains identical outputs.
- [ ] Measure actual runtime before making any time-to-reproduce claim.

Publication gate

- [ ] Report all attempted specifications or register the confirmatory one in advance.
- [ ] Distinguish failure of a project threshold from failure of a theory.
- [ ] Release code and non-restricted derived artifacts.
- [ ] Obtain an independent statistical review.

- [ ] Submit the corrected study as a replication attempt only after the above items pass.

Archived output names

The September 2025 audit trail identifies these primary outputs:

- `results/dop_dunnhumby_beauty_spec_q90_b2_m20.csv`
- `results/dop_instacart_shampoo_specification_compliant.csv`
- `results/dj_bodycare.csv`
- `results/moderation_bodycare.csv`
- `results/cep_coverage_complete.csv`
- `results/dirichlet_bodycare.csv`

These filenames identify the historical snapshot. They are not evidence that the current pipeline reproduces from a clean checkout.

References

Marketing-science foundations

1. Ehrenberg, A. S. C. (1959). "The Pattern of Consumer Purchases." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 8(1), 26-41. https://doi.org/10.2307/2985810
2. Goodhardt, G. J., Ehrenberg, A. S. C., & Chatfield, C. (1984). "The Dirichlet: A Comprehensive Model of Buying Behaviour." *Journal of the Royal Statistical Society: Series A*, 147(5), 621-643. https://doi.org/10.2307/2981696
3. Ehrenberg, A. S. C., Uncles, M. D., & Goodhardt, G. J. (2004). "Understanding Brand Performance Measures: Using Dirichlet Benchmarks." *Journal of Business Research*, 57(12), 1307-1325. https://doi.org/10.1016/j.jbusres.2002.11.001

These sources define empirical regularities and model structure. They do not supply the project-specific MAD<=0.015 or Pearson r>=0.80 gates.

Data sources

1. Chen, D. (2012). *Online Retail II*. UCI Machine Learning Repository. https://doi.org/10.24432/C5CG6D
2. dunnhumby. *The Complete Journey*, Source Files. https://www.dunnhumby.com/source-files/
3. Instacart. *Instacart Online Grocery Shopping Dataset 2017*. Kaggle. https://www.kaggle.com/datasets/instacart/instacart-online-grocery-shopping-2017
4. Ni, J., Li, J., & McAuley, J. (2019). "Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects." *Proceedings of EMNLP-IJCNLP 2019*, 188-197. https://doi.org/10.18653/v1/D19-1018
5. McAuley Lab, UC San Diego. *Amazon Review Data (2018)*. https://cseweb.ucsd.edu/~jmcauley/datasets/amazon_v2/

Archived project evidence

The numerical values in this report were checked against the September 2025 JSONL logs and CSV files named in the reproduction checklist. The logs for the dunnhumby, Instacart, and Amazon runs include input-hash prefixes. The UCI logs do not include a recoverable input hash or source commit.

Legal & Compliance

This report is provided for informational and educational purposes. It does not constitute professional, investment, or marketing advice. The archived analyses have not been peer reviewed and should not be used to make budget, targeting, or portfolio decisions.

The source datasets remain subject to their respective licenses and access terms. This report does not redistribute restricted raw data.

Contact information submitted through the distribution form is retained for up to 12 months for document delivery and research updates. Deletion requests may be sent to support@visageaiconsulting.com.