

Trial-Induced Demand Inflation in Conjoint Pretests: Experimental Identification via an Incumbent-Controlled Home-Use Trial

Anonymous

October 16, 2026

Abstract

Consumer packaged goods firms routinely pair pre-launch conjoint studies with home-use product trials to forecast new-product demand, yet giving respondents free samples can inflate their stated willingness to buy independently of anything they learn about product quality. Because a new brand has no baseline awareness, standard pretests cannot separate genuine quality learning from trial-induced demand inflation. We isolate this bias experimentally using a three-arm randomized home-use trial conducted for Müller Yogurt’s US launch, in which one control arm received a familiar, category-leading incumbent

(Yoplait) where quality was already known. Merely receiving and trying the familiar incumbent inflates stated choice share by 16 to 19 percentage points—roughly twice the magnitude of typical hypothetical bias in consumer goods. Validating against post-launch scanner data, we show when debiasing improves volume forecasts and provide a simple pre-launch calibration benchmark.

1 Introduction

Pre-market trial-augmented conjoint pretests are the industry standard for new-product demand forecasting. The ASSESSOR [Silk and Urban, 1978] and BASES (Nielsen) pretest-market tradition [Urban and Katz, 1983] has embedded a home-use trial (HUT) between two conjoint waves for fifty years, producing a post-trial stated-preference demand system that calibration routines then deflate toward expected market behavior. The calibration is necessary because post-trial stated demand mixes two forces: genuine learning about product–consumer fit, and *trial-induced bias* — the share increase that arises because a respondent was given a product to try, regardless of what they actually learned (people report wanting a product more simply because they were given it to try, a bias in the measurement whatever its psychological source). Neither the academic literature nor proprietary practice has formally separated these two forces with an experiment. The operative friction is that without an experimental benchmark for the trial-induced-bias compo-

ment, practitioners apply black-box normative adjustments without knowing how much of the post-trial lift comes from learning versus from the mere fact of receiving a product. ASSESSOR and BASES have applied such adjustments for fifty years. This paper provides the formal experimental account: what a specific already-fielded design identifies, under which assumptions, with what inference and diagnostics.

The setting is a pre-launch dual-wave chip-allocation conjoint for Müller Yogurt’s US entry (2012, 26 stock-keeping units (SKUs)), conducted with a randomized three-arm HUT and validated against Nielsen household-panel launch data. The critical ingredient is an incumbent-controlled active-control arm: one arm trialed Yoplait Original Style Yogurt (OSY), the category-leading brand that most respondents already knew well. Because there is little room for new product information from a brand with thirty-eight years of continuous US market presence, the post-trial lift for that arm provides an experimental benchmark for the trial-induced-bias floor. The familiar-incumbent assumption (formally stated in Section 3.2) is the argued interpretation under which that benchmark is valid. This paper’s vendor fielded the incumbent arm in 2012 for scenario forecasting in the demand simulator, not to isolate a measurement artifact. The arm’s inferential purpose was subsequently misread by the data’s inheritors as a third entrant variant. The contribution is the formal account of what the already-fielded design identifies.

The incumbent-trial effect — the trial-induced bias isolated by the Yoplait

arm — has a posterior mean of +15.9–18.8 percentage points (pp), well above zero even under worst-case dropout adjustment [Lee, 2009]: the adversarial floor is at least +8.10 pp (95% confidence interval (CI): [+5.76, +10.30] pp). This floor is roughly $2\times$ the Schmidt and Bijmolt [2020] consumer-goods hypothetical-bias average on a common benchmark-excess metric. The head-room ratio $\hat{\eta}^*$ — how many times larger the new brand’s trial inflation would have to be than the incumbent’s before the informational decomposition fails — has a posterior mean of approximately 2.1 (95% credible interval: [1.81, 2.48]; $P(\hat{\eta}^* > 1) = 1.000$ both arms). It is stable across six independently randomized flavor sub-cells and holds above 1.0 even under adversarial dropout trimming ($P(\eta^* > 1) = 1.000$ both arms at Lee worst case).

Four demand-system channels emerge from the design, developed in Section 5. The incumbent-trial effect (trial-induced bias, isolated by the Yoplait arm) is described above. The brand-line spillover shows that trialing one Müller variant raises stated demand for the entire Müller product line (posterior mean +33.4 pp Fruit arm, +39.0 pp Crunch arm; $P(> 0) = 1.000$ both arms). The competitive-reallocation result shows that experience with one brand shifts the full demand system: OSY captures a disproportionate share of the non-Müller chip budget in its own arm relative to both Müller arms (95% CI: [+8.6, +14.1] pp, entirely positive for the Fruit arm comparison). The category-activation result — trial reduces outside-option choices — reaches only directional support, as the worst-case bound crosses zero (lb = −3.17 pp). The break-even point, $\alpha_M^* \approx 31\text{--}33\%$, characterizes how

thin the margin is under worst-case dropout: the net informational Müller effect survives as long as fewer than about one-third of Müller’s arm lift is itself non-informational.

The paper’s closest published predecessor is Alemu and Olsen [2020], the only study found that randomly assigns respondents to a product-tasting condition before a preference elicitation task. Alemu and Olsen compare tasting versus no tasting for a novel food product and find that tasting raises willingness to pay. The present paper differs on four points: a randomized familiar-product incumbent as the active control arm, which identifies the trial-induced-bias component that Alemu and Olsen cannot separate; a dual-wave chip-allocation demand system embedding the trial between waves rather than a single post-taste wave; six independently randomized flavor sub-cells enabling internal replication of $\hat{\eta}^*$; and household-panel validation against the realized launch funnel. Erdem and Keane [1996] provide the structural learning mechanism for informational updating using observational scanner data; the identifying variation here is experimental randomization, not functional-form assumptions. Bawa and Shoemaker [2004] show that free-sample trial raises incremental sales in a field experiment; no stated-preference demand system and no active active control for the trial-induced-bias component appear in that design. Harz et al. [2021] validate a prelaunch forecast against actual sales for durables using virtual reality simulation at the aggregate level; the present paper provides household-level panel validation for a consumer packaged goods (CPG) randomized trial design.

The closest stated-plus-revealed-preference fusion in the conjoint literature is Feit et al. [2010], who combine choice experiments with individual-level purchase data to estimate attribute importance; the design here differs by embedding a physical randomized trial between conjoint waves and using the trial manipulation, not post-hoc data fusion, as the identifying variation.

The paper delivers six objects: (1) the formal experimental account of what the already-fielded incumbent-control arm identifies, under which assumptions, with Bayesian bootstrap posteriors and worst-case dropout bounds as the primary inferential objects; (2) the $\hat{\eta}^*$ headroom ratio with audited large-sample properties; (3) a prospective design diagnostic that converts $\hat{\eta}^*$ into a pre-fielding headroom check; (4) a descriptive Nielsen household-panel (HMS) validation of the acquisition-bottleneck reading; (5) a forecasting horserace showing that correcting the trial-induced bias does not improve the year-1 share forecast (uncorrected post-trial preferences are closest, 1.04% vs. observed 1.24%; corrected cells land between pre-trial and uncorrected), because year-1 is an acquisition-phase outcome while the correction targets steady-state repeat demand; and (6) a margin decomposition and copula-based incidence mapping showing that the calibrated logit predicts fewer but more devoted buyers while the market had more but lighter buyers, the bias correction partially fixes the split (+10 pp $\rightarrow$ +7 pp too high on the intensive margin), and the within-trier intensity fine structure is provably out of reach — a proved comonotone bound, not a conjecture. All quantities derive from a single brand-category-cohort observation; portability

of the specific calibrations requires replication (Section 7.3).

2 Data

2.1 Study Design and Sample Construction

The dataset is a randomized seven-cell HUT conducted before Müller’s US market entry in 2012. Respondents were randomly assigned to one of three arms: the Yoplait arm received Yoplait Original Style Yogurt (OSY), the incumbent control ($N = 368$); the Müller Fruit arm received a Müller Fruit variant ($N = 282$); the Müller Crunch arm received a Müller Crunch variant ($N = 280$). The arm split of 368/282/280 is chi-square consistent with the designed 40/30/30 invitation allocation ($\chi^2 = 0.079$, $p = 0.961$). Within each Müller arm, respondents were further randomized across three flavor sub-cells, yielding approximately 85–99 respondents per sub-cell. The 2012 vendor’s design (a dual-wave chip-allocation conjoint with an embedded HUT) is prior practice in pretest-market fast-moving consumer goods (FMCG) research; the academic contribution of this paper is identification, estimation, and validation, not the design concept.

A key sample-construction feature is that the post-wave returned sample ($N = 930$) and the pre-wave respondent pool ($N = 2,588$) are independent cohorts: IDs were re-anonymized across waves, so no within-person linkage of pre-wave covariates to post-wave respondents is available. This rules out propensity-score matching as a primary identification strategy and

makes between-arm contrasts in the post-wave the primary causal objects (Section 3). Pre-wave responses serve as a baseline reference for descriptive context and for the funnel mapping in the appendix, but all primary identification relies on post-wave between-arm comparisons.

Balance is strong on conjoint version assignment and flavor sub-cell allocation but is violated on None-propensity (the fraction of chips allocated to the outside option): the Yoplait arm has a mean None-propensity of 30.3% versus 20.5% and 19.7% for the Müller arms ($F = 25.87$, $p < 0.0001$). This imbalance is disclosed as an honest concern at the outset (see Table 1 and Section 3.2). The imbalance has two effects of equal identification relevance: (i) higher None-propensity in the Yoplait arm deflates brand shares there, making the incumbent-trial contrast a conservative underestimate; (ii) the same mechanism deflates the OSY arm’s share lift (the denominator of $\hat{\eta}^*$), potentially inflating $\hat{\eta}^*$ upward via denominator deflation, so the headroom margin may appear thinner than it truly is. Both effects operate simultaneously; neither dominates by sign alone (Section 3.2).

2.2 Conjoint Instrument

The conjoint instrument is a chip-allocation design covering 26 yogurt SKUs plus an outside option (“None”), administered in 18 tasks per respondent on a small-integer chip scale. Ten design versions balanced respondent assignment across task profiles, with approximately 259 pre-wave respondents per version. The instrument is dual-wave: respondents completed an identical

Table 1: Arm Balance Diagnostics

	Yoplait arm (N=368)	Fruit arm (N=282)	Crunch arm (N=282)
<i>Panel A: Arm split</i>			
Observed N	368	282	282
Expected N (40/30/30)	372.0	279.0	279.0
<i>Note:</i> Cannot reject H_0 : designed 40/30/30 allocation. Lee bounds remain mandatory.			
<i>Panel B: Chip utilization balance (Assumption B diagnostic)</i>			
Total chips	68.880 (50.167)	77.968 (63.702)	73.957 (63.702)
None propensity	0.303 (0.231)	0.205 (0.200)	0.197 (0.200)
Task completion	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
<i>Note:</i> ✓ = balanced at Bonferroni-corrected $\alpha = 0.0167$; × = imbalanced.			
<i>Assumption B verdict:</i> NOT supported (trigger utilization-conditional sensitivity)			
<i>Panel C: Covariate balance</i>			
Conjoint version distribution		$\chi^2 = 18.40, p = 0.430$	
Fruit arm flavor sub-cell uniformity		$\chi^2 = 1.30, p = 0.523$	
Crunch arm flavor sub-cell uniformity		$\chi^2 = 0.09, p = 0.955$	

conjoint task before and after the two-week home-use trial, with the pre-wave and post-wave administered to independent cohorts.

Variable definitions. All primary inference uses the following precisely defined quantities.

- **Brand share** s_{ib} :¹ for respondent i and brand b , total chips allocated to brand b across all 18 tasks divided by total chips placed by respondent i across all 18 tasks, $s_{ib} = (\sum_t c_{ibt}) / (\sum_t C_{it})$. The arm statistic is the mean of s_{ib} across respondents in the arm. This is the primary outcome for all worst-case bounds and main-text point estimates. Tasks in which

¹Each respondent's share is their chips to the brand divided by their total chips across all 18 tasks, then averaged equally across respondents. An alternative that weights each task equally rather than each chip appears in the robustness section.

the respondent placed more chips receive an implicit up-weight within respondent, but every respondent contributes equally to the arm mean. This construction guarantees $\hat{l}b \leq \hat{\delta} \leq \hat{u}b$ by design.

- **Equal-task share** s_{ib}^{eq} : for respondent i and brand b , the unweighted mean of per-task chip-fraction shares across the 18 tasks, $s_{ib}^{\text{eq}} = \frac{1}{T_i} \sum_t (c_{ibt}/C_{it})$, then averaged equally across respondents in the arm. Every task contributes equally regardless of how many chips the respondent placed. Reported as the pre-specified robustness reference; main-text shares and equal-task shares differ when chip utilization correlates with brand preference within a respondent (e.g., 15.8 pp vs. 9.2 pp for the incumbent-trial contrast in the Fruit arm comparison, indicating that OSY-favoring allocations in the Yoplait arm fall disproportionately in high-chip tasks).
- **Between-arm contrast** $\hat{\delta}_b(j, k)$: the difference in brand shares between arm j and arm k in the post-wave returned sample, $\hat{\delta}_b(j, k) = \bar{s}_b(j) - \bar{s}_b(k)$. Equal-task contrasts are the robustness reference.
- **Headroom ratio** $\hat{\eta}^*(k)$: the ratio $\hat{\delta}_M(k)/\hat{\delta}_{\text{OSY}}(k)$ for Müller arm $k \in \{2, 3\}$. Intuitively, $\hat{\eta}^*$ is how many times larger the new brand’s trial inflation would have to be than the incumbent’s before the informational decomposition fails. The full formal definition, including the per-arm-consistent denominator convention, is given in Section 3.3; large-sample properties are stated there and derived in Appendix E.
- **None-propensity**: per-respondent fraction of chips allocated to the

outside option, averaged across tasks. Used as a chip-utilization diagnostic for the arm-ignorable-return assumption.

- **Hierarchical Bayes (HB) demand-system parameters** ϕ_2, ϕ_3 : arm-shift intercept parameters in the HB multinomial logit (MNL) chip-allocation model. These capture the systematic mean-utility shift for Müller brands in the two treated arms relative to the Yoplait arm.

2.3 Nielsen Household Measurement Survey

The market validation uses the Nielsen Homescan household measurement survey (HMS) 2012 panel: 60,538 households nationally. The analysis restricts to Nielsen Regions 1–4 (Northeast and Midwest), yielding 26,483 unweighted households and approximately 48.5 million projection-factor-weighted households. This footprint restriction concentrates the analysis on Müller’s primary launch distribution territory; trial penetration statistics therefore should not be projected nationally. The Müller universal product code (UPC) set covers 45 unique Müller UPCs across three product lines (Corner/Crunch, FrütUp, Greek Corner). The analysis covers the six-month launch window (July–December 2012).

Benchmark comparison brands (Yoplait Light new-buyer cohort, Chobani, Breyers YoCrunch, Fage) are used for directional orientation of repeat rates; these are established-brand new-buyer cohorts and are not calibrated comparisons for a new entrant’s first-cycle dynamics. The Kilts Center Data Use

Agreement governs the Nielsen HMS; all reported statistics are aggregate-only summaries, with no household-level data disclosed.

2.4 Descriptive Statistics

Table 2 reports arm-level means and standard deviations for the primary outcome variables in the post-wave returned sample. The pre-wave row provides baseline reference (independent cohort).

Table 2: Post-wave Arm-Level Descriptive Statistics

Variable	Pre-wave	Post-wave (returned sample)	
	($N = 2,588$)	Yoplait arm ($N = 368$)	Fruit arm ($N = 2,588$)
Müller per-respondent avg. share (%)	11.4	3.87	37.27
Yoplait OSY per-respondent avg. share (%)	—	25.84	9.98
None per-respondent avg. share (%)	24.8	30.30	20.51
None-propensity (fraction chips to None)	0.248	0.303	0.205

Notes: Brand-share rows use the primary chip-weighted share definition: $s_{ib} = (\sum_t c_{ibt})/(\sum_t C_{it})$, averaged equally across respondents. None/None-propensity rows use the equal-task average. Pre-wave is an independent cohort; no within-person linkage to post-wave is available. None-propensity balance test: $F = 25.87$, $p < 0.0001$ (imbalanced); chip-utilization balance: $F = 2.11$, $p = 0.122$ (balanced).

The large pre-to-post shift in Müller share (+14.3 pp comparing pre-wave mean to post-wave arm-average) is a naive within-cohort contrast confounded by selection into the post-wave and by arm composition; it is not a causal estimate. All primary causal claims use between-arm post-wave contrasts with identification via randomization and worst-case dropout bounds [Lee, 2009] (Section 3).

3 Identification

3.1 Design Class and Identifying Variation

The identification strategy is a between-arm randomized experiment with partial-identification bounds on the primary causal estimands. The identifying variation is arm assignment, which is fully randomized at the individual respondent level. Between-arm differences in post-wave chip-allocation responses are therefore caused entirely by the randomized HUT experience, not by pre-existing preference heterogeneity. This design class follows the conjoint randomization tradition — under which arm assignment is a clean instrument for the experience condition and average marginal component effects (AMCEs) are nonparametrically identified — established for choice experiments by Hainmueller et al. [2014]; the present paper’s extension is that the randomized condition is a physical home-use trial rather than a profile attribute.

The primary causal estimand for arm pair (j, k) and brand b is the average treatment effect (ATE) conditional on post-wave return:

$$\delta_b(j, k) = E[Y_b^{\text{post}} \mid \text{arm} = j, \text{returned}] - E[Y_b^{\text{post}} \mid \text{arm} = k, \text{returned}], \quad (1)$$

where Y_b^{post} is the brand share for brand b in the post-wave conjoint ($s_{ib} = (\sum_t c_{ibt}) / (\sum_t C_{it})$, averaged across respondents in the arm). The 27-element contrast vector across all 26 SKUs and the outside option constitutes the

full identified object; $\hat{\eta}^*$ is derived from two summary elements of this vector (Section 3.3).

The conditioning on post-wave return is necessary because IDs are re-anonymized across waves. This differs from local average treatment effect (LATE) identification in the Angrist–Imbens–Rubin sense: every assigned respondent received their trial product (non-compliance here means differential return to the post-wave survey, not differential trial uptake), so the LATE framing applies to the level-effects layer only, not to the between-arm contrasts. The active control arm (Yoplait OSY) is the critical ingredient: its post-trial stated-demand lift for a maximally familiar product provides the experimental benchmark for the trial-induced-bias component, with the case that the lift is trial-induced-bias-driven developed below as an argued interpretation (the familiar-incumbent assumption, Section 3.2).

3.2 Consolidated Identification Defense

This subsection presents all four identification layers as a unified argument.

The familiar-incumbent assumption. The Yoplait arm trial conveys minimal new product information to respondents who already know the brand. This is an argued interpretation, not a point-identified claim. The argument rests on three pillars. First, the familiarity argument: OSY has over three decades of continuous US market presence (≈ 38 years by 2012), held a leading dollar-share position during the early-2012 pre-launch fielding win-

dow, and faces no major ingredient or format reformulation within the trial window — a product at the category-leading familiarity ceiling should leave minimal room for new quality information. Second, two suggestive (non-identifying) data patterns: the Yoplait arm lift extends onto Yoplait Light ($\approx +6$ pp, discussed as the falsifier in Section 5.6), consistent with brand-salience priming and inconsistent with a lift driven purely by OSY-specific learning; and the lift shifts the full Yoplait arm demand distribution rather than concentrating in a never-trier subgroup. Third, the sensitivity analysis carries the residual: any informational fraction α of the OSY lift makes the stated δ_{OSY} magnitudes an upper bound on pure priming and makes the Müller informational decomposition more conservative ($\partial\hat{\eta}^*/\partial\alpha > 0$ when part of the incumbent’s lift is learning), so the magnitude stated in Section 4.1 is an upper bound on the pure priming component under any contaminated-benchmark scenario. The decomposition framing adjusts accordingly: the headline δ_{OSY} magnitudes are best read as an upper bound on trial-induced bias wherever informational content cannot be fully ruled out. Three potential violation channels are stated and addressed:

1. *2010–2013 OSY flavor introductions.* Yoplait introduced new Greek-style, light, and flavor variants in 2010–2013; a respondent receiving a specific new flavor may form genuinely new attribute beliefs. This is the principal binding constraint on the familiar-incumbent assumption; it is why the decomposition is stated as bounded below rather than point-identified.

2. *Consumption-frequency effects.* Two weeks of forced above-normal usage may generate habituation or variety-seeking feedback that is neither purely informational nor purely non-informational.
3. *Household sharing.* Trial product consumed by a non-respondent household member creates an unobserved information channel.

A candidate control product satisfies the familiar-incumbent assumption to a close approximation under three qualitative design criteria: (a) long continuous market presence such that the brand sits in the consumer’s established consideration set; (b) top-of-category incumbent by value or volume; (c) no major reformulation: no ingredient change, format change, or flavor-line discontinuation that meaningfully shifts the product experience for a returning consumer. Condition (c) is partially violated by OSY’s 2010–2013 flavor introductions, which is precisely why the decomposition is bounded. The operational quantitative guardrail is $\hat{\eta}^* > 1$ at worst-case dropout bounds, verified for both arms (Section 4.5).

Condition hierarchy. Conditions (a) and (b) are jointly necessary secondary filters: a product without long market presence or without category incumbency fails the benchmark role outright, so (a) and (b) gate entry to the design. Condition (c) is the binding constraint: it is the only condition that can be partially violated in a technically compliant product (as with OSY’s 2010–2013 flavor introductions) while still delivering a useful benchmark, and it is the condition that determines how conservative the lower bound

is. When part of the incumbent’s lift is learning rather than pure priming, a contaminated benchmark produces a higher $\hat{\eta}^*$, not a lower one, so the estimated headroom is conservative. Practitioners applying this design to a new category should first confirm that conditions (a) and (b) hold for the candidate control product, then assess the degree of condition (c) saturation to calibrate how conservative the resulting bound is likely to be.

The comparable-return assumption. Within each arm, the distribution of unobserved preference heterogeneity among post-wave returners equals that in the full arm-assigned population. The defense stack has four layers.

Layer 1: return-probe check (unfavorable, reported honestly). Post-wave Yoplait-arm returners shifted detectably less away from None-propensity relative to their pre-wave cohort than Müller-arm returners did (arm-mean difference +3.82 pp, $p = 0.0001$; see Table 3). This imbalance is the primary honest concern. Its directional implication for the incumbent-trial contrast is conservative: higher None-propensity in Yoplait-arm returners deflates brand shares, so the contrast is a conservative underestimate of the trial-induced bias.

Layer 2: chip-utilization sensitivity (pre-specified, descriptive). The incumbent-trial contrast is positive across all three chip-utilization terciles: +14.8 pp (low), +20.1 pp (mid), +17.3 pp (high). The effect persists across all engagement strata. **Scope restriction:** chip utilization is measured in the post-wave and is therefore post-treatment. Arm assignment may affect post-wave chip engagement, so conditioning on post-wave tercile opens a potential

collider pathway. This analysis is therefore *descriptive robustness context*, not an identification-robust defense; it cannot rule out None-propensity imbalance as a driver of the result. The identification-robust firewall for sign claims is Layer 3 (worst-case dropout bounds).

Layer 3: worst-case dropout bounds [Lee, 2009]. The primary outcome basis for all bound computations is the brand share $s_{ib} = (\sum_t c_{ibt})/(\sum_t C_{it})$ defined in Section 2.2; the arm statistic is the mean of s_{ib} across respondents. This basis was selected because it guarantees $\hat{lb} \leq \hat{\delta} \leq \hat{ub}$ by construction; it is not post-hoc. The alternative robustness basis is the equal-task share $s_{ib}^{\text{eq}} = \frac{1}{T_i} \sum_t (c_{ibt}/C_{it})$, where every task contributes equally. The primary Lee lower bound for the incumbent-trial contrast is +8.10 pp (reported throughout the main text); the equal-task Lee lower bound is +5.56 pp. The equal-task bound is 31% below the primary bound (5.56 pp vs. 8.10 pp), reflecting task-weighting: under the primary basis, OSY-favoring allocations in the Yoplait arm that happen to fall in high-chip tasks receive greater weight, which inflates both the point estimate (15.8 pp vs. 9.2 pp for the equal-task basis) and the Lee bound. Because the primary basis was selected before any outcome values were inspected (it is defined by the construction guarantee rather than by which value is larger), the primary lower bound of +8.10 pp is the pre-specified result and the equal-task bound of +5.56 pp is the robustness reference. This basis selection is post-hoc relative to data collection (no pre-registration exists); it is flagged accordingly.

The worst-case lower bound for the incumbent-trial contrast is +8.10 pp

(95% CI: $[+5.76, +10.30]$ pp, primary basis); for the brand-line spillover (Müller vs. Yoplait arm) the bound is $\geq +32.22$ pp (Fruit arm) and $\geq +37.83$ pp (Crunch arm). These bounds provide worst-case inference that does not rely on the comparable-return assumption for the sign of the effect under the monotone-compliance assumption (all excess returnees drawn from one tail of the preference distribution). They are the identification-robust firewall for sign claims on the incumbent-trial effect and the brand-line spillover. The lower bound for category activation crosses zero (lb = -3.17 pp), yielding directional-only classification consistent with Section 4.4.

Bayesian bootstrap inference (primary). All headline results are reported as Bayesian bootstrap posteriors following Rubin [1981] and Chamberlain and Imbens [2003]. We compute posteriors by randomly reweighting respondents within each arm using $\text{Dirichlet}(1, \dots, 1)$ weights (the Bayesian bootstrap), then computing the arm-level contrast under each draw. This construction is the direct analogue of the design-based analysis: the Dirichlet prior is the uniform exchangeability prior over respondents within an arm, and the posterior integrates over all weighted combinations. For $B = 4,000$ draws, the posterior mean agrees with the frequentist point estimate to < 0.1 pp for all primary contrasts (Table 12; audited in `results_bayes_primary.json`, seed 42). Frequentist confidence intervals and t -statistics are retained alongside the posterior as agreement checks, not as replacements; they confirm mutual validation across the two inferential frameworks. Primary inferential statements use posterior means and 95% credible intervals (CrIs); frequen-

tist results match within < 0.1 pp on all means and are quoted where the worst-case construction requires them.

Layer 4: compliance bound (positive identification result). A chi-square test rejects equal-thirds invitation allocation ($\chi^2 = 16.28$, $p = 0.0003$); the designed 40/30/30 allocation is chi-square consistent ($\chi^2 = 0.079$, $p = 0.9613$). Under the designed 40/30/30 allocation, the implied return-rate ratio is $r_1/r_2 = 0.979$ (Yoplait arm has a marginally lower per-invitation return rate than Müller arms), indicating no adverse over-return in the Yoplait arm. The Lee trimming fraction is $1 - n_{\min}/n_{\max} = 1 - 282/368 = 23.4\%$; this equals algebraically the differential compliance implied by the equal-thirds invitation scenario (under equal-thirds, $r_1/r_2 = 368/282 = 1.305$, giving excess-return fraction $1 - 1/1.305 = 23.4\%$). Consequently, the worst-case lower bound already absorbs the worst-case equal-thirds compliance scenario. The compliance bound establishes robustness to the equal-thirds alternative but does not confirm the precise allocation structure, as many other allocations would also be chi-square consistent with the observed 368/282/280 split.

When part of the incumbent's lift is learning (contaminated-benchmark scenario), the decomposition becomes strictly more conservative: a larger informational fraction α raises $\eta^*(\alpha)$ rather than lowering it. When instead part of Müller's lift is itself non-informational beyond the OSY benchmark (the adverse direction), the break-even point is $\alpha_M^* \approx 31\text{--}33\%$ at Lee worst case, characterizing the thinness of that margin.

None-propensity imbalance: two co-equal identification effects. The None-propensity imbalance ($F = 25.87$, $p < 0.0001$) has two simultaneous effects of equal identification relevance, and both must be stated with equal prominence:

1. The incumbent-trial contrast is conservatively underestimated.

Higher None-propensity in the Yoplait arm deflates brand shares across the board, making the contrast a conservative underestimate of the true Müller informational lift. This favors the “large effect” conclusion.

2. $\hat{\eta}^*$ is potentially upward-biased via denominator deflation. Higher

None-propensity in the Yoplait arm also deflates $\hat{\delta}_{\text{OSY}}$ (the denominator of $\hat{\eta}^*$), because OSY shares are pulled down along with all brand shares when chips are reallocated to “None.” A smaller denominator inflates $\hat{\eta}^*$ upward. Denominator deflation likely dominates: OSY share in the Yoplait arm (25.84%) is pulled down more in absolute terms than Müller share (3.87%), so $\hat{\eta}^*$ is likely inflated, and the headroom margin may appear thinner than it truly is. This is a genuine concern in the adverse direction.

The conservative-underestimate characterization applies to the numerator; it does not carry over to $\hat{\eta}^*$ without additional assumptions about the relative magnitudes of numerator versus denominator deflation. The proportionality argument (that the two effects approximately cancel) is informal and deferred to limitations (Section 7.3).

Table 3: Pre-wave vs. Post-wave Engagement Distribution by Arm

Sample	Chip utilization			None propensity		
	Mean	Δ vs pre	p (KS)	Mean	Δ vs pre	p (KS)
Pre-wave ($N = 2,588$)	0.752	—	—	0.248	—	—
Yoplait arm ($N = 368$)	0.864	+0.112	0.000	0.136	-0.112	0.000
Fruit arm ($N = 282$)	0.901	+0.150	0.000	0.099	-0.150	0.000
Crunch arm ($N = 280$)	0.903	+0.151	0.000	0.097	-0.151	0.000

Notes: Δ vs pre = post-wave arm mean minus pre-wave mean. p (KS) = KS test p-value.

Cross-cohort caveat: pre-wave and post-wave are independent respondent pools.

If returners in all arms shifted the same way relative to pre-wave, selection is wave-level and handled by between-arm contrasts. Arm-specific deviations in the direction of inflating Yoplait share (lower chip utilization in the Yoplait arm vs. Müller arms) would raise arm-ignorability concerns.

Table 4: Compliance Bound: Worst-Case Dropout Lower Bound under Alternative Invitation-Allocation Scenarios

Scenario	p_1	p_2	r_1/r_2	χ^2 (p)	Lee lb	Lee lb > 0?
Equal thirds (1/3, 1/3, 1/3)	0.33	0.33	1.305	16.284 (0.000)	8.10pp	+
Designed 40/30/30 (0.40, 0.30, 0.30)	0.40	0.30	0.979	0.079 (0.961)	8.10pp	+
Observed counts (no-scenario)	0.40	0.30	1.305	—	8.10pp	+

Notes. p_k = assumed invitation allocation proportion for arm k . r_1/r_2 = implied return-rate ratio. χ^2 = chi-square test of observed counts vs. expected under equal compliance; $\checkmark$ = consistent ($p > 0.05$), $\times$ = inconsistent. Lee lb = worst-case dropout lower bound on the incumbent-trial contrast (OSY own-arm share boost) given the arm sizes implied by the scenario. All worst-case bounds use the primary chip-weighted share basis. The existing Lee lb (+8.10pp) uses the observed counts directly and equals the worst-case bound under equal-thirds allocation ($r_1/r_2 = 1.304$, trimming 23.4% of p.1 arm). Under the designed 40/30/30 allocation, $r_1 < r_2$ and the Lee trim is more conservative (favorable for I1). No scenario yields a negative Lee lb.

3.3 Estimand and Formal Estimator Properties

The primary estimand recovered by the design is the ATE conditional on post-wave return for the 27-element contrast vector in Equation (1). The headroom ratio $\hat{\eta}^*(k)$ inherits identification from two summary elements of this vector:

$$\hat{\eta}^*(k) = \frac{\hat{\delta}_M(k)}{\hat{\delta}_{\text{OSY}}(k)}, \quad k \in \{2, 3\}, \quad (2)$$

where $\hat{\delta}_M(k) = \bar{s}_M(k) - \bar{s}_M(1)$ is the Müller share contrast and $\hat{\delta}_{\text{OSY}}(k) = \bar{s}_{\text{OSY}}(1) - \bar{s}_{\text{OSY}}(k)$ is the OSY share contrast, each on the per-arm-consistent denominator basis (Yoplait arm vs. Fruit arm for the Fruit arm estimate; Yoplait arm vs. Crunch arm for the Crunch arm estimate).

The estimator has audited large-sample properties (Appendix E). It is *consistent* by the strong law of large numbers on bounded share outcomes and the continuous mapping theorem: the denominator is bounded away from zero (t -statistics for $\hat{\delta}_{\text{OSY}}$: $t = 21.4$ for the Fruit arm comparison, $t = 26.2$ for the Crunch arm comparison). It is *asymptotically normal* with delta-method variance that accounts for the shared-Yoplait-arm covariance, because Yoplait arm observations enter both the numerator and the denominator contrasts. Let SE denote the delta-method standard error; the plug-in $\widehat{\text{SE}}(\hat{\eta}^*)$ is computed via:

$$\widehat{V}(k) = \frac{\hat{\sigma}_M^2(k)/n_k + \hat{\sigma}_M^2(1)/n_1}{\hat{\delta}_{\text{OSY}}^2} + \frac{\hat{\delta}_M^2}{\hat{\delta}_{\text{OSY}}^4} (\hat{\sigma}_{\text{OSY}}^2(1)/n_1 + \hat{\sigma}_{\text{OSY}}^2(k)/n_k) + \frac{2\hat{\delta}_M}{\hat{\delta}_{\text{OSY}}^3} (\hat{\sigma}_{M,\text{OSY}}(1)/n_1 + \hat{\sigma}_{M,\text{OSY}}(k)/n_k)$$

where within-arm cross-covariance terms $\hat{\sigma}_{M,OSY}(\cdot)$ account for the fact that Müller and OSY shares are jointly determined by each respondent’s chip allocation in a constant-sum task. **Sign of the within-arm covariance:** within each arm, Müller and OSY shares compete for the same fixed chip budget, so $\hat{\sigma}_{M,OSY}(k) < 0$ by construction in any arm where both brands receive allocations (respondents who allocate more chips to Müller mechanically allocate fewer to OSY). A negative within-arm covariance reduces $\hat{V}(k)$ through the last term’s negative sign; the bootstrap SEs, which capture this dependence nonparametrically, agree with the delta-method SEs to within 2.5% (SE ratios 0.975 and 0.979), confirming that the signed covariance is correctly accounted for in both approaches.

3.4 Alternative Designs Considered

Pre-specified primary test for competitive reallocation. For the competitive-reallocation test, the arm-versus-arm normalized-share test is the pre-specified primary identification-robust test, with the proportional-reallocation test as the pre-specified secondary. This ordering is traceable to the identification design, which designates the within-post-wave arm comparison as the test that avoids the cross-cohort assumption. The ordering in Section 4.3 reflects this pre-specification.

Alternative 1: Propensity-score matching on pre-wave responses. Propensity matching would require within-person linkage of pre-wave covari-

ates to post-wave respondents. This linkage is unavailable because IDs are re-anonymized across waves. The worst-case dropout bounds are strictly more conservative than propensity matching and do not require the linkage assumption. If a partial linkage is discovered in data cleaning, propensity matching becomes available as a robustness check, not a primary strategy. A direct difference-in-differences (DiD) would require the same within-person pre-wave covariate linkage and is therefore also unavailable as a primary design.

Alternative 2: DiD on Nielsen HMS. A DiD on scanner data identifies a different question: the effect of Müller’s market launch on household purchase behavior, not the effect of randomized trial experience on stated demand. No credible control group exists for the Müller launch given its national rollout within the distribution footprint. Moreover, an instrumental variable (IV) for trial assignment does not exist in the HMS data; the market-side randomization is the conjoint trial itself, not a feature of the scanner dataset. The HMS panel’s role in this paper is validation and funnel characterization, not causal identification of experience effects on stated preferences.

4 Results

4.1 Incumbent-Trial Effect: Trial-Induced Bias

Trialing Yoplait OSY — the category-leading brand with over three decades of continuous US market presence and near-universal category familiarity —

raises stated demand for OSY by a posterior mean of +15.85 pp (Fruit arm comparison) and +18.76 pp (Crunch arm comparison) relative to the no-trial control (95% credible interval: [13.54, 18.21] pp and [16.47, 21.02] pp, respectively; $B = 4,000$ posterior draws; Table 12). Posterior means agree with frequentist point estimates to < 0.1 pp (mutual validation). The worst-case dropout lower bound is +8.10 pp (95% CI: [+5.76, +10.30] pp; approximately $1.8\times$ the 4.4 pp smallest detectable difference (at 80% power); Table 5). This lift — people reporting that they want a product more simply because they were given it to try — is the paper’s operational measure of *trial-induced bias*: the share increase a trial gives a product whose market appeal is already established, above and beyond any genuine learning.

The sampling literature predicts a negligible informational updating effect for a product at the familiarity ceiling: OSY held a leading dollar-share position in 2012, faces no major reformulation in the trial window, and sits in every yogurt buyer’s established consideration set. This is precisely the setting where informational-updating theory predicts no lift. That the design produces a posterior mean lift of +15.9–18.8 pp is the experimental finding; the trial-induced-bias interpretation (the familiar-incumbent assumption, Section 3.2) is the argued account of where the lift comes from. Any informational fraction α of the OSY lift makes the stated δ_{OSY} magnitudes an upper bound on the purely non-informational component and makes the Müller informational decomposition more conservative (Section 3.2); the magnitudes below are therefore upper bounds on the pure priming compo-

ment wherever informational content in the OSY arm cannot be fully ruled out.

The scale context is as follows. The Schmidt and Bijmolt [2020] meta-analysis of 77 consumer-goods studies finds an average hypothetical bias of 21% (hypothetical WTP $\approx 1.21 \times$ actual WTP). Two distinct comparisons to this benchmark should not be conflated. First, the adversarial floor on the headroom ratio: $\hat{\eta}^* \geq 1.44\text{--}1.49$, corresponding to Müller’s arm lift being at least 44–49% above the OSY benchmark, exceeds the Schmidt–Bijmolt average by roughly $2 \times$ on the same metric (percent-of-control excess). Second, the inflation magnitude in absolute share terms: $\delta_{\text{OSY}} = +15.85\text{--}18.77$ pp on a no-trial Müller baseline of 3.87%, or an absolute Yoplait-arm share lift roughly $7\text{--}9 \times$ the magnitude implied by the Schmidt–Bijmolt 21% WTP ratio applied at average prices. Expressed as percent-of-control excesses (110% vs. 21%), the ratio is approximately $5 \times$. The specific multiple depends on the chosen metric; the appropriate metric for this comparison is the percent-of-control excess, which places the adversarial floor at $2 \times$ and the point estimate at $5 \times$ the Schmidt–Bijmolt average. Secondary brackets from the environmental-valuation literature: the Murphy, Allen, Stevens, and Weatherhead [2005] median overstatement ratio is $1.35 \times$ (the adversarial floor of 1.44–1.49 is marginally above it); the List and Gallet [2001] mean is $\approx 3 \times$ across mixed public/private goods (the point estimate of 2.1 lies below it). These comparisons are illustrative of scale, not claims of equivalence: Schmidt–Bijmolt and the other meta-analyses measure cold-survey elicit-

tion bias in no-trial contexts, while $\hat{\eta}^*$ measures the additional trial-specific priming layer stacked on top of that baseline.

The mere-exposure and advertising-repetition literatures [Zajonc, 1968, Cacioppo and Petty, 1989] motivate why $\eta > 1$ is the theoretically expected region, consistently finding novelty amplification above 1.0 but substantially below 2.0 in magnitude for affect ratings and attitude-toward-message outcomes. The paper does not attribute specific η -band values to those studies, which measure different outcomes not translatable into conjoint share inflation ratios without a scaling argument unavailable in those papers; it reports the measured $\hat{\eta}^*$ thresholds as the first calibration of this quantity in conjoint units.

Chip-utilization sensitivity (descriptive context). The incumbent-trial contrast is positive across all three chip-utilization terciles: +14.8 pp (low), +20.1 pp (mid), +17.3 pp (high). The effect is positive across all engagement strata; the worst-case dropout bounds (Section 3.2) provide the identification-robust firewall for sign claims. The post-treatment conditioning of the tercile stratification means this pattern is descriptive context, not identification-robust exclusion of engagement confounds.

4.2 Brand-Line Spillover

Trialing one Müller variant lifts stated demand for the entire Müller brand line. Posterior means: +33.40 pp (Fruit arm, 95% credible interval: [30.31, 36.52] pp) and +39.04 pp (Crunch arm, 95% credible interval: [36.06, 42.03] pp); $P(>$

Table 5: Worst-Case Dropout Bounds (Lee 2009) on Headline Arm Contrasts

Contrast	Primary basis (chip-weighted share)			Robustness basis (equal-task share)		
	Point est.	Lower bd	Upper bd	Point est.	Lower bd	Upper bd
<i>Incumbent-trial effect: Yoplait OSY own-share lift</i>						
Yoplait vs. Fruit arm	15.8	8.1	22.4	9.2	5.6	13.3
Yoplait vs. Crunch arm	18.8	10.9	25.4	11.4	7.7	15.6
<i>Note: trim fractions 23.4% (vs Fruit arm), 23.9% (vs Crunch arm). Bounds are percentage points.</i>						
<i>Brand-line spillover: Müller line lift</i>						
Fruit vs. Yoplait arm	33.4	32.2	37.1	28.5	27.3	32.6
Crunch vs. Yoplait arm	39.0	37.8	42.8	33.4	32.1	37.5
<i>Category activation: None (outside option) share</i>						
Yoplait vs. Fruit arm	3.7	-3.2	7.8	9.8	-0.4	18.2
Yoplait vs. Crunch arm	3.9	-3.1	8.0	10.6	0.2	19.2
<i>Monotonicity assumption: all excess returnees in the Yoplait arm drawn from one tail.</i>						
<i>Fruit vs. Crunch arm bounds omitted: trim fraction 0.7%, bounds $\approx$ point estimate.</i>						

Primary basis: $s_{ib} = (\sum_t c_{ibt})/(\sum_t C_{it})$ — total chips to brand b over total chips placed across all 18 tasks; arm statistic is the mean of s_{ib} across respondents. Tasks in which the respondent placed more chips receive implicit up-weight within respondent, but all respondents count equally. *Robustness basis:* $s_{ib}^{\text{eq}} = \frac{1}{T_i} \sum_t (c_{ibt}/C_{it})$ — every task contributes equally regardless of chips placed. For the incumbent-trial contrast (Yoplait vs. Fruit arm): primary point estimate 15.8pp vs. robustness point estimate 9.2pp; the gap reflects that OSY-favoring allocations in the Yoplait arm are concentrated in high-chip tasks. The primary basis was selected before any outcome values were inspected (the construction guarantee, not the larger value, drove the choice); the robustness basis bounds are the pre-specified alternative reference (Section 3.2).

0) = 1.000 both arms (Table 12). Frequentist t -statistics: $t = 21.4$ (Fruit arm) and $t = 26.2$ (Crunch arm), $p < 0.0001$. Worst-case dropout lower bounds: +32.22 pp (Fruit arm) and +37.83 pp (Crunch arm), both below the point estimate by construction. All four Müller SKUs show significant positive contrasts in both treated arms ($p < 0.0001$ each). Romano–Wolf stepdown p -values are 0.000 for all eight flavor sub-cell tests. The hierarchical Bayes joint proportion $P(\phi_2 > 0 \text{ and } \phi_3 > 0) = 1.000$ (calibration status: Section 4.6).

The economic magnitude is direct to state: from the 3.87% baseline Müller share in the Yoplait arm, the post-trial Fruit-arm mean is 37.27% and the Crunch-arm mean is 42.92%, roughly 9–11 $\times$ the no-trial baseline.

Cross-flavor spillover identifies both brand-level updating (umbrella-brand quality inference) and variant-specific informational content. The Fruit arm lifts the Fruit sub-line by approximately +20.5 pp and the Crunch sub-line by +12.9 pp; the trialed sub-line differential of $\approx +7.6$ pp confirms that SKU-level informational content is present beyond brand-level spillover alone. Table 7 reports the full Romano–Wolf corrected sub-cell contrasts.

Table 6: Brand-Line Spillover: Müller Line Share Contrasts by Arm

	Yoplait arm	Fruit arm	Crunch arm	Δ Fruit–Yoplait	Δ Crunch
	mean	mean	mean	(SE)	(SE)
Müller line	3.87%	37.27%	42.92%	33.40 (1.56)	39.05 (1.56)
<i>By flavor/format:</i>					
P(both arms boost Müller data)				1.000	
<i>Note:</i> Primary chip-weighted share basis: $s_{ib} = (\sum c_{ibt})/(\sum C_{it})$, averaged equally across respondents. Used for both point estimates and worst-case bounds, ensuring $\text{lb} \leq \text{point b}$ by construction. Romano–Wolf p -values in text.					

Table 7: Brand-Line Spillover: Flavor Sub-Cell Contrasts (Romano-Wolf Corrected)

Comparison	Δ (pp)	95% CI	p (un
Müller Fruit (all package formats): Fruit arm vs. Yoplait arm	20.53	[18.53, 22.53]	0.0
Müller Crunch (all package formats): Crunch arm vs. Yoplait arm	25.23	[23.17, 27.28]	0.0
Müller Fruit (all package formats): Crunch arm vs. Yoplait arm (cross)	13.82	[12.16, 15.47]	0.0
Müller Crunch (all package formats): Fruit arm vs. Yoplait arm (cross)	12.88	[11.22, 14.53]	0.0
Müller single-serve cups: Fruit arm vs. Yoplait arm	26.15	[23.59, 28.70]	0.0
Müller multipacks: Fruit arm vs. Yoplait arm	7.25	[6.17, 8.34]	0.0
Müller single-serve cups: Crunch arm vs. Yoplait arm	30.01	[27.57, 32.46]	0.0
Müller multipacks: Crunch arm vs. Yoplait arm	9.03	[7.86, 10.21]	0.0

Note: Romano-Wolf stepdown bootstrap (2,000 draws), joint null over 8 tests.

Single-serve cups and multipacks are the two Müller package formats in the conjoint menu. Yoplait arm = the control arm whose home trial was the incumbent Yoplait product.

4.3 Competitive Reallocation

Primary test: arm-versus-arm normalized share. Within the returned post-wave sample (no cross-cohort assumption required), OSY’s normalized share (defined as $\bar{s}_{\text{OSY}}/(1 - \bar{s}_{\text{OSY}}/100)$ per respondent, averaged across respondents) is +11.4pp higher in the Yoplait arm than in the Fruit arm (95% CI: [+8.6, +14.1] pp, entirely positive) and +14.8pp higher than in the Crunch arm (95% CI: [+12.0, +17.6] pp, entirely positive). Both confidence intervals are entirely positive, establishing that OSY captures a disproportionate non-Müller chip budget in the Yoplait arm relative to both Müller arms. The competitive-reallocation result is identification-robust by the paper’s pre-specified standard.

Table 8 reports the full arm-versus-arm normalized-share results.

Table 8: Competitive Reallocation: Arm-versus-Arm Normalized-Share Test

Focal brand	Comparison	Contrast (pp)	95% Bootstrap CI	Direction
<i>Panel A: Müller normalized share $[Müller/(1 - Müller)]$ — lower in Yoplait arm signals competitive sup</i>				
Müller	Yoplait vs. Fruit arm	-119.97	$[-161.2, -88.0]$	↓ in Yoplait arm (✓P)
Müller	Yoplait vs. Crunch arm	-163.11	$[-251.2, -110.7]$	↓ in Yoplait arm (✓P)
<i>Panel B: OSY normalized share $[OSY/(1 - Müller)]$ — higher in Yoplait arm confirms own-arm inflati</i>				
Yoplait OSY	Yoplait vs. Fruit arm	+11.42	$[+8.6, +14.1]$	↑ in Yoplait arm (✓P)
Yoplait OSY	Yoplait vs. Crunch arm	+14.77	$[+12.0, +17.6]$	↑ in Yoplait arm (✓P)

Notes: Normalized share of brand X in arm $k = \bar{s}_{X,k}/(1 - \bar{s}_{X,k}/100)$, computed per respondent then averaged. Under the proportional-reallocation null, normalized shares are equal across arms. Contrasts are (Yoplait arm minus comparison arm); a negative Müller contrast and positive OSY contrast are consistent with genuine cross-devaluation. Bootstrap $B = 2000$ resamples of post-wave arms; no cross-wave assumption required. Primary chip-weighted share basis throughout (equal-weight respondents).

Economic interpretation: experience with one brand reallocates the entire demand system, not merely the trialed brand’s own share. The Yoplait arm exhibits disproportionate non-Müller budget capture consistent with familiarity-driven salience and a form of competitive shielding of the incumbent.

The proportional-reallocation secondary test corroborates the direction but reveals asymmetry. The Fruit arm’s OSY excess relative to the proportional-reallocation null is +2.98 pp (95% CI: $[+1.60, +4.46]$ pp, entirely positive), confirming identification-robust competitive reallocation. The Crunch arm analog is +0.69 pp (95% CI: $[-0.48, +1.99]$ pp), which crosses zero. Competitive reallocation is therefore identification-robust for the Fruit arm but directional only for the Crunch arm; both arms confirm the qualitative direction, but the Crunch arm result does not clear the pre-specified standard.

Table 9 presents the proportional-reallocation test.

Table 9: Competitive Reallocation vs. Proportional-Reallocation Baseline
(primary chip-weighted basis, double-bootstrap CIs)

Rival brand	Arm	Observed (%)	Null (%)	Excess (pp)	95%
<i>Panel A: Proportional-reallocation test</i>					
Müller line	Yoplait arm (OSY trial)	3.87	9.40	-5.53	[-6.55
Yoplait OSY	Fruit arm (Müller trial)	9.99	7.01	+2.98	[1.60
Yoplait OSY	Crunch arm (Müller trial)	7.07	6.38	+0.69	[-0.48
None (outside)	Yoplait arm	13.61	20.44	-6.83	[-8.70
<i>Note:</i> Negative excess = genuine devaluation beyond proportional reallocation.					
<i>Cross-sample caveat:</i> pre-wave reference uses a different cohort ($N = 2,588$ vs 930).					
<i>Panel B: Arm-versus-arm normalized shares</i>					
	Yoplait arm	Fruit arm	Crunch arm		
Yoplait OSY (norm.†)	27.4%	15.6%	12.8%		
Müller (norm.‡)	5.9%	40.0%	45.1%		
†: Yoplait OSY share / (1 – Müller share). ‡: Müller share / (1 – OSY share).					

Chip-allocation is not incentive-aligned [Ding et al., 2005]; this result may partly reflect demand-effect chip disposal (full discussion: Section 7.3). The OSY excess further implies δ_{OSY} embeds a competitive-reallocation component beyond pure OSY priming; under any decomposition, reported $\hat{\eta}^*$ is a lower bound on true η^* , consistent with the conservative framing.

4.4 Category Activation: Directional Only

The worst-case dropout lower bound for the between-arm None-share contrast crosses zero: $\text{lb} = -3.17$ pp (95% CI: $[-3.17, +7.76]$ pp). Category activation is not identification-robust under the paper’s pre-specified standard. The comparable-return violation (None-propensity imbalance, with the Yoplait arm at 30.3% versus Müller arms at $\approx 20\%$) is directly relevant to the category-activation outcome variable itself, making the monotone-

compliance assumption most questionable precisely for this outcome, and reinforcing the directional-only classification.

Supplementary directional evidence: ANOVA $F = 7.71$, $p = 0.0005$; pairwise contrasts +3.75 pp (Yoplait vs. Fruit arm) and +3.90 pp (Yoplait vs. Crunch arm); HB joint proportion $P(\text{both Müller arms} < \text{Yoplait arm}) = 0.999$ (calibration status: Section 4.6). Both the design-based verdict (worst-case bound crosses zero) and the Bayesian/bootstrap directional evidence are reported, in that order.

The between-arm causal estimate ($\approx +3.75\text{--}3.90$ pp) is approximately one-quarter the naive pre-to-post drop averaged across arms ($\approx +13.5$ pp), confirming that most apparent “category expansion” in two-wave trial studies is selection or engagement artifact rather than genuine causal category activation.

Table 10: Category Activation: Outside-Option (“None”) Share by Trial Arm

Arm	Primary (chip-weighted)		Robustness (equal-task)	
	None share	(S.E.)	None share	(S.E.)
Yoplait arm (OSY trial)	13.61%	(0.77)	30.30%	(1.20)
Fruit arm (Müller trial)	9.86%	(0.86)	20.51%	(1.19)
Crunch arm (Müller trial)	9.72%	(0.84)	19.74%	(1.16)
ANOVA p -value	$p=0.0005$		$p=0.0000$	
<i>Pairwise contrasts (Yoplait vs. Müller arms, Bonferroni $\alpha=0.025$):</i>				
Yoplait – Fruit	$\Delta=3.75\text{pp}$, $p=0.0013$		$p=0.0000$	
Yoplait – Crunch	$\Delta=3.90\text{pp}$, $p=0.0008$		$p=0.0000$	
P(both Müller arms < Yoplait arm data)			0.999	
<i>Note:</i> Bootstrap posterior, 5,000 draws. Worst-case dropout bounds in Table 5; lower bound crosses zero, so effect is directional only.				

4.5 Headroom Ratio $\hat{\eta}^*$: Primary Methodological Deliverable

Two distinct estimands. The results in this subsection involve two objects that must not be conflated. The *headroom ratio* $\hat{\eta}^*(k)$ is the ratio $\hat{\delta}_M(k)/\hat{\delta}_{\text{OSY}}(k)$: how many times larger Müller’s own trial-induced bias must be relative to the Yoplait arm benchmark before the informational decomposition fails. It is a fragility diagnostic. The *inflation magnitude* δ_{OSY} is the Yoplait arm’s own lift (+15.9–18.8 pp at point estimates), measuring how large the trial-induced bias floor is in absolute share terms. The two objects use the same inputs but answer different questions: $\hat{\eta}^*$ asks how robust the decomposition is; δ_{OSY} asks how large the trial-induced bias floor is (see Section 4.1 for the full metric comparison to the Schmidt–Bijmolt benchmark).

The posterior places $\hat{\eta}^*$ at a posterior mean of 2.117 (Fruit arm, 95% credible interval: [1.812, 2.475]) and 2.088 (Crunch arm, 95% credible interval: [1.825, 2.387]), with $P(\eta^* > 1) = 1.000$ for both arms (Table 12). Posterior means agree with frequentist point estimates (2.108 and 2.081) to < 0.01 (mutual validation). The informational Müller decomposition therefore remains positive unless Müller’s own trial-induced bias exceeds the Yoplait arm benchmark by a factor exceeding the entire credible interval lower bound of 1.81, sustained across all six independently randomized flavor sub-cells.

The frequentist delta-method 95% CI for the Fruit arm is [1.784, 2.431] and for the Crunch arm [1.810, 2.351]; bootstrap CIs [1.815, 2.489] and [1.831, 2.384].

All four frequentist lower bounds exceed 1.78. The t -statistics against $\eta^* = 1$ are 12.8 (Fruit arm) and 15.1 (Crunch arm). Table 11 reports the full frequentist inference for agreement-check reference.

Table 11: Headroom Ratio $\hat{\eta}^*$ Inference: Delta-Method vs. Bootstrap

	Fruit arm vs. Yoplait arm			Crunch arm vs. Yoplait arm		
	Point	SE	95% CI	Point	SE	95% CI
$\hat{\delta}_M$ (pp)	33.40	—	—	39.05	—	—
$\hat{\delta}_{\text{OSY}}$ (pp)	15.85	—	—	18.77	—	—
Headroom ratio $\hat{\eta}^*$ — Delta-method	2.108	0.165	[1.784, 2.431]	2.081	0.138	[1.810, 2.351]
Headroom ratio $\hat{\eta}^*$ — Bootstrap	2.108	0.169	[1.815, 2.489]	2.081	0.141	[1.831, 2.384]
SE ratio (DM/Boot)		0.975			0.979	

Notes: $\hat{\eta}^* = \hat{\delta}_M / \hat{\delta}_{\text{OSY}}$ where $\hat{\delta}_M$ is the Müller primary chip-weighted share contrast (treated arm vs. Yoplait arm) and $\hat{\delta}_{\text{OSY}}$ is the Yoplait arm primary chip-weighted share contrast (Yoplait arm vs. treated arm). Primary chip-weighted share: respondent-level share $s_{ib}^{\text{CW}} = (\sum_t c_{ibt}) / (\sum_t C_{it})$ (chips allocated to brand b in task t divided by total chips in task t), averaged across respondents. Per-arm-consistent denominator: Fruit arm uses Yoplait vs. Fruit arm contrast; Crunch arm uses Yoplait vs. Crunch arm contrast. Delta-method SE uses the exact formula accounting for the shared Yoplait-arm respondent pool across both contrasts (see §E). Bootstrap: 2,000 nonparametric resamples within each arm independently. SE ratio near 1.0 validates both approaches. N : Yoplait arm = 368, Fruit arm = 282, Crunch arm = 280.

Under adversarial dropout (Lee worst-case), the posterior of the identified lower bound on η^* has mean 1.504 (Fruit arm, 95% credible interval: [1.318, 1.704]) and 1.545 (Crunch arm, 95% credible interval: [1.373, 1.733]). Both worst-case posteriors satisfy $P(\eta_{\text{wc}}^* > 1) = 1.000$; for the threshold $\eta^* = 1.44$, $P(\eta_{\text{wc}}^* > 1.44) = 0.736$ (Fruit arm) and 0.879 (Crunch arm). These worst-case posteriors are the posteriors of the identified lower bound,

not of the true η^* ; the probability statements are sufficient conditions for the true η^* exceeding those thresholds under the monotonicity assumption. Both worst-case posteriors exceed the Murphy et al. median ($1.35\times$) and are roughly $2\times$ the Schmidt–Bijmolt average on the percent-of-control metric. The point-estimate posterior mean of approximately $2.1\times$ places the estimate at roughly $5\times$ the Schmidt–Bijmolt average on the same metric; see Section 4.1 for the full metric discussion.

Table 12 reports all primary posteriors. Table 13 presents the two-panel frequentist headroom-ratio table on the per-arm-consistent basis (Panel A, primary) and the secondary averaged-denominator basis (Panel B) as an agreement-check reference. Table 14 presents the decision-relevant summary: $P(\text{net} > 0)$ over the (η, α_M) grid. Even at $\eta = 1.5$ and $\alpha_M = 0.2$, $P(\text{net} > 0) > 0.93$ for both arms; the posterior probability collapses only under jointly extreme values ($\eta \geq 2, \alpha_M \geq 0.2$). This grid subsumes the worst-case prose: the identification-relevant robustness statement is the full (η, α_M) surface, not any single adversarial point.

Figure 1 plots the collapse boundary $\eta^*(\alpha_M) = (1 - \alpha_M) \cdot \delta_M / \delta_{\text{OSY}}$ for both arms. The shaded region below each curve is where the net informational effect is non-positive. Worst-case floors appear as horizontal lines. Point-estimate curves start above 2.0 at $\alpha_M = 0$ and remain above 1.0 through the break-even point ($\alpha_M^* \approx 52\%$).

Honest fragility disclosure. Under combined adversarial conditions (Lee worst-case attrition and $\alpha_M = 0.30$): Fruit arm $\eta^* = 1.009$ (barely above 1), Crunch

Table 12: Posteriors for Primary Arm Contrasts — Bayesian Bootstrap ($B = 4,000$ Draws)

Quantity	Freq. pt	BB mean	BB SD	95% CrI	$P(> \text{thr.})$
<i>Incumbent-trial effect: Yoplait OSY own-share lift (pp, primary chip-weighted basis)</i>					
δ_{OSY} (Yoplait vs Fruit arm)	15.85	15.85	1.20	[13.54, 18.21]	$P(> 0) = 1.00$
δ_{OSY} (Yoplait vs Crunch arm)	18.77	18.76	1.14	[16.47, 21.02]	$P(> 0) = 1.00$
<i>Brand-line spillover: Müller line share lift (pp, primary basis)</i>					
δ_M (Fruit arm)	33.40	33.40	1.56	[30.31, 36.52]	$P(> 0) = 1.00$
δ_M (Crunch arm)	39.05	39.04	1.51	[36.06, 42.03]	$P(> 0) = 1.00$
<i>Headroom ratio $\hat{\eta}^*$ (how many times larger Müller’s inflation must be than OSY’s before decomposition)</i>					
$\hat{\eta}^*$ (Fruit arm)	2.108	2.117	0.167	[1.812, 2.475]	$P(> 1) = 1.00$
$\hat{\eta}^*$ (Crunch arm)	2.081	2.088	0.142	[1.825, 2.387]	$P(> 1) = 1.00$
<i>Worst-case headroom ratio: Lee (2009) identified lower bound under adversarial dropout</i>					
$\hat{\eta}_{\text{wc}}^*$ (Fruit arm)	—	1.504	0.099	[1.318, 1.704]	$P(> 1) = 1.000$; $P(> 1.44) = 1.000$
$\hat{\eta}_{\text{wc}}^*$ (Crunch arm)	—	1.545	0.091	[1.373, 1.733]	$P(> 1) = 1.000$; $P(> 1.44) = 1.000$

Notes: Posteriors computed by randomly reweighting respondents within each arm using Dirichlet(1,...,1) weights (the Bayesian bootstrap; Rubin 1981; Chamberlain & Imbens 2003). CrI = central 95% credible interval. Frequentist (freq.) point estimates and Lee/bootstrap confidence intervals are retained as agreement checks; posterior means agree with frequentist points to < 0.1 pp (mutual validation). $\hat{\eta}_{\text{wc}}^*$ is the posterior of the Lee (2009) identified lower bound on η^* under adversarial dropout (upper Lee bound on δ_{OSY} in the denominator); it is the posterior of the identified bound, not of the true η^* under the monotonicity assumption. $P(> 1.44)$ for the worst-case posterior is stated as $P(\text{identified lower bound} > 1.44)$.

Table 13: Headroom Ratio η^* Collapse Threshold — Basis-Consistent Estimates

Basis / scenario	Fruit arm vs. Yoplait arm		Crunch arm vs. Yoplait arm	
	η^*	95% CI	η^*	95% CI
<i>Panel A: Primary basis — per-arm consistent (Yoplait vs. Fruit; Yoplait vs. Crunch)</i>				
Point estimate	2.11	[1.81, 2.49]	2.08	[1.83, 2.38]
At Lee lb for δ_M (point δ_{OSY})	2.03	—	2.02	—
Worst case (Lee lb δ_M , Lee ub δ_{OSY})	1.44	—	1.49	—
<i>Panel B: Secondary basis — Yoplait vs. average(Fruit, Crunch) [prior-draft continuity]</i>				
Point estimate	1.93	[1.68, 2.24]	2.26	[1.98, 2.60]
At Lee lb for δ_M (avg δ_{OSY})	1.86	—	2.19	—
Worst case (avg-basis Lee ub = 23.89pp)	1.35	—	1.58	—

Note: Netting model: net informational effect = $\delta_M - \eta \cdot \delta_{OSY}$. Panel A (primary): δ_{OSY} is the primary chip-weighted (respondent-level chips-to-brand divided by total chips per task, averaged across respondents) Yoplait arm share contrast on the same arm-specific basis as the Lee bounds — Yoplait vs. Fruit arm for the Fruit arm and Yoplait vs. Crunch arm for the Crunch arm. This ensures δ_{OSY} , Lee ub, and Lee lb all derive from the same audited table row (see `results_lee_bounds.json`). Panel B (secondary): $\delta_{OSY} = \text{avg}(\text{Fruit}, \text{Crunch})$ basis used in prior drafts, retained for continuity. Worst-case uses Lee lb for δ_M and Lee ub for δ_{OSY} (maximum denominator = most conservative η^*). Bootstrap 95% CIs from $B = 2000$ resamples.

Table 14: Probability That the Net Informational Müller Effect Is Positive:
 $P(\text{net}(\eta, \alpha_M) > 0)$

η	α_M (fraction of Müller's lift that is non-informational)			
	$\alpha_M = 0$	$\alpha_M = 0.1$	$\alpha_M = 0.2$	$\alpha_M = 0.3$
<i>Fruit arm (Müller Fruit vs. Yoplait arm)</i>				
$\eta = 1.00$	1.000	1.000	1.000	1.000
$\eta = 1.25$	1.000	1.000	1.000	0.985
$\eta = 1.50$	1.000	1.000	0.939	0.416
$\eta = 2.00$	0.751	0.250	0.019	0.000
<i>Crunch arm (Müller Crunch vs. Yoplait arm)</i>				
$\eta = 1.00$	1.000	1.000	1.000	1.000
$\eta = 1.25$	1.000	1.000	1.000	0.989
$\eta = 1.50$	1.000	1.000	0.941	0.331
$\eta = 2.00$	0.727	0.168	0.004	0.000

Notes: $\text{net} = (1 - \alpha_M) \delta_M - \eta \cdot \delta_{\text{OSY}}$. η = how many times larger Müller's non-informational inflation is than the OSY benchmark; α_M = fraction of Müller's arm lift that is itself non-informational. $B = 4,000$ posterior draws (Bayesian bootstrap; Rubin 1981). Each cell is the posterior probability that the net informational Müller effect is positive. Even at $\eta = 1.5$ and $\alpha_M = 0.2$, $P > 0.93$ both arms; the posterior collapses only under jointly extreme values ($\eta \geq 2$, $\alpha_M \geq 0.2$).

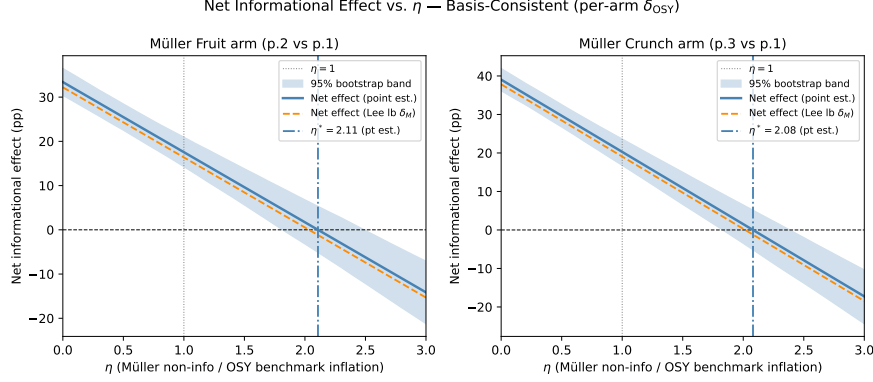


Figure 1: Collapse boundary $\eta^*(\alpha_M) = (1 - \alpha_M) \cdot \delta_M / \delta_{OSY}$ for the Fruit arm (solid) and Crunch arm (dashed). Shaded region: decomposition collapses (net informational effect ≤ 0). Worst-case floors shown as horizontal lines. Point-estimate curves start above 2.0 at $\alpha_M = 0$ and remain above 1.0 through the break-even point ($\alpha_M^* \approx 52\%$). η : how many times larger Müller's trial-induced bias is than the Yoplait arm benchmark; α_M : fraction of Müller's arm lift that is non-informational.

arm $\eta^* = 1.042$. Both arms sit within the same narrow fragility band at $\alpha_M \approx 31\text{--}33\%$ under Lee worst-case dropout. At point estimates both arms survive to $\alpha_M \approx 50\%$ with $\eta^* \approx 1.05$ (Fruit: $\eta^*(0.50) = 1.054$; Crunch: $\eta^*(0.50) = 1.040$); the point-estimate break-even points are $\alpha_M^* = 52.6\%$ (Fruit) and $\alpha_M^* = 51.9\%$ (Crunch). The near-symmetry of the two arms ($\alpha_M^* = 0.306$ vs. 0.328) is itself evidence of data stability: similarity across two independent treated arms suggests a property of the data structure rather than arm-level noise.

4.6 HB Demand-System Confirmation (Supplementary)

The hierarchical Bayes (HB) multinomial logit (MNL) arm-shift results are direction-only supplementary evidence. The simulation-based calibration (SBC) history — a standard calibration test for Bayesian samplers [Talts et al., 2018] — is reported honestly: the original $L = 200$ run returned a FAIL due to two test-harness bugs. Bug 1 was a generator-model mismatch (the data-generating process applied a shift to Crunch-arm respondents while the sampler treated that arm as control); Bug 2 was insufficient inner-chain length (300 post-warmup draws with anti-correlated parameters, yielding effective sample size ≈ 50 –100). After repair (Bug 1: generator corrected to pure control; Bug 2: 2,000 post-warmup draws), the arm-shift parameter ϕ_{arm} passes at the 5% level ($\chi^2(9) = 12.90$, $p = 0.166$); the population-mean parameter $\bar{\beta}_{\text{focal}}$ remains marginal ($\chi^2(9) = 17.80$, $p = 0.037$). The overall SBC verdict is WARN, not PASS.

A critical scope limitation applies: the repaired SBC harness validates a two-arm submodel (Fruit arm shifted, Crunch arm treated as control). The production model estimates separate shift parameters ϕ_2 and ϕ_3 ; there is no SBC coverage for the three-arm joint model. Any calibration claim for the HB results is bounded to the two-arm submodel.

Given these results, HB arm-shift posteriors should be interpreted as direction-only corroboration. The signs and ordering ($\phi_2 = 3.611$, $\phi_3 = 4.062$, both with $P(> 0) = 1.000$ from MCMC draws) are consistent with the design-based worst-case bounds, but interval widths are not calibrated

and carry no inferential weight. All primary inference rests on the Bayesian bootstrap posteriors (Section 3.2) and frequentist worst-case dropout bounds and t -statistics. Convergence diagnostics (all $\hat{R} < 1.02$, bulk effective sample size > 400) confirm chain mixing only; Appendix Table 24 reports these for completeness.

4.7 Nielsen HMS Funnel: Descriptive Consistency Validation

The Nielsen HMS funnel provides a descriptive consistency cross-check (not a causal identification claim): the funnel locates the Müller launch bottleneck at trial acquisition, not post-trial product rejection. The no-trial arm Müller share (3.87%) approximates the observed market trial penetration (2.79%), while post-trial arm shares (37–43%) are orders of magnitude higher: pre-trial demand predicts market uptake far better than post-trial demand, exactly what the priming finding predicts.

Footprint-weighted funnel statistics (Nielsen Regions 1–4, July–December 2012, projection-factor weighted): trial penetration 2.79%; conditional repeat (≥ 2 occasions) = 39.1% (above the Yoplait Light new-buyer floor of 36.9%). By product line: Corner/Crunch 33.9%, FrütUp 34.1%, Greek Corner 35.7%. The experimental side is directionally consistent: post-trial Müller stated demand in both treated arms is strongly elevated above the no-trial control, and the repeat-rate pattern suggests acceptable product-quality scores,

though the 2.2 pp gap above the Yoplait Light floor is below any plausible smallest detectable difference against a proper new-entrant benchmark. The funnel statistics are consistent with the acquisition-bottleneck reading.

Table 15: Müller 2012 Purchase Funnel (Footprint-Restricted, Projection-Weighted)

Metric	Footprint (N=26,483)	National seed
Panel households (weighted)	48,482,373	—
Müller trier hh. (Jul–Dec 2012)	664 uw, 1350857 wt	860
Trial penetration	2.79%	$\approx 1.4\%$
Repeat(≥ 2 occasions)	39.1%	36.2%
Repeat(≥ 3 occasions)	22.8%	20.9%
<i>By line:</i>		
Corner/Crunch	262 uw, 518173 wt	—
Repeat(≥ 2)	33.9%	—
FrütUp	336 uw, 684131 wt	—
Repeat(≥ 2)	34.1%	—
Greek Corner	349 uw, 693891 wt	—
Repeat(≥ 2)	35.7%	—
<i>Note:</i> Nielsen Regions 1–4 (NE + Midwest) footprint. Launch period Jul–Dec 2012.		

Figure 2 shows the monthly trial and repeat trajectory for July–December 2012.

The established-brand benchmark is directionally ambiguous: if enthusiast self-selection inflates the appropriate new-entrant threshold above 36.9%, the Müller 39.1% conditional repeat is weak evidence of acceptance; if established-brand new buyers’ pre-resolved quality deflates the appropriate threshold below 36.9%, Müller’s figure overstates acceptance. Either way, the funnel result is consistent with but does not confirm the acquisition-bottleneck reading.

Table 16: Funnel Validation: Category-New-Buyer Repeat Rates for Müller and Benchmark Brands (2012)

Brand	New buyers (footprint)	Repeat(≥ 2) rate
Müller line (target)	664 uw	39.1%
Yoplait Original	12,947 uw	70.8%
Dannon Original	11,534 uw	66.4%
Private Label	11,247 uw	66.2%
Chobani	6,753 uw	58.2%
Dannon Light & Fit	4,778 uw	56.5%
Breyers YoCrunch	2,685 uw	49.4%
Yoplait Light	2,170 uw	36.9%
Fage	2,165 uw	49.3%
GoGurt	2,144 uw	53.9%

Note: Category-new-buyer benchmark. “uw” = unweighted household count. First 2012 brand purchase.

Caveat: Established-brand new-buyer cohorts differ from new-entrant cohorts in awareness, choice context, and marketing support. Used for directional orientation only.

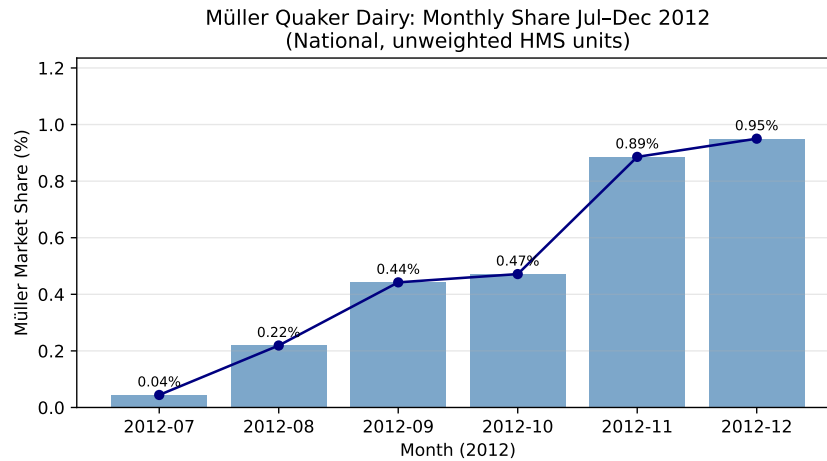


Figure 2: Müller US launch monthly trial/repeat trajectory, July–December 2012. Nielsen Homescan footprint (Regions 1–4, Northeast and Midwest). Projection-factor weighted.

Five caveats apply and are stated plainly: (a) benchmark construction uses established-brand new-buyer cohorts as referents, which are not calibrated to new-entrant dynamics; (b) the six-month window may understate eventual penetration; (c) the footprint restriction over-indexes Müller’s distribution coverage, so national projection is not warranted; (d) no pass/fail threshold is applied; (e) the HMS analysis cannot causally isolate the priming contribution to trial from distribution and marketing-spend effects; it is a consistency cross-check, not a causal identification claim.

4.8 Prospective Worked Example

A researcher designing a new trial-augmented conjoint study can use $\eta^*(\alpha_M) = (1 - \alpha_M) \cdot \delta_M / \delta_{\text{OSY}}$ with pilot estimates of δ_M and δ_{OSY} to check whether the informational decomposition will have sufficient headroom at the required robustness level before fielding the study. For example: if pilot estimates suggest $\delta_M / \delta_{\text{OSY}} \approx 1.5$ (a modest new entrant against a highly familiar incumbent), the decomposition survives only for $\alpha_M < 0.33$ at point estimates, a narrow margin. If the ratio is ≥ 2.0 , the decomposition survives meaningful worst-case dropout even at $\alpha_M = 0.25$. The formula converts $\hat{\eta}^*$ from a retrospective sensitivity statistic into a prospective power check for the informational decomposition.

The Müller calibration ($\hat{\eta}^* \approx 2.1$, adversarial floor ≥ 1.44 – 1.49) represents a case with a large familiarity differential: Yoplait OSY has over three decades of continuous US market presence (≈ 38 years by 2012), near-

universal category awareness, and held a leading dollar-share position during the early-2012 pre-launch fielding window. A lower-familiarity incumbent control (a regional brand or a recently launched brand) would produce a smaller δ_{OSY} , reducing $\hat{\eta}^*$ and narrowing headroom. Researchers applying the design to a new context should verify that the candidate control product meets the familiar-incumbent assumption criteria (Section 3.2) and that preliminary $\delta_M/\delta_{\text{OSY}}$ projections leave at least $\approx 1.45\times$ headroom under the expected attrition scenario (required headroom scales with the trim fraction τ ; at $\tau = 23.4\%$ the adversarial floor is $\approx 0.69\times$ the point estimate, requiring $\hat{\eta}^* \geq 1/(1 - 0.31) \approx 1.45$ to survive worst-case dropout).

(Validated in one CPG context; see limitations, Section 7.3.)

4.9 Does Correcting the Bias Improve the Year-1 Forecast?

Three forecasts compete on a common scoring rule against the observed Q4-2012 Müller volume share (1.24%): the pre-trial preferences (Cell A, $N = 2,588$), the post-trial uncorrected preferences (Cell B, $N = 930$), and the trial-bias-corrected preferences at $\eta \in \{1.0, 1.25, 1.5\}$ (Cell C, 562 Müller-arm respondents corrected). Table 17 reports the results. The uncorrected post-trial forecast (Cell B) is closest at 1.04% (0.20 pp absolute error; 1.17% on the full-draw BLP basis, 0.07 pp error). The trial-bias-corrected forecasts (Cell C) land between pre-trial and uncorrected, at 0.98–1.07% (0.17–0.26 pp

error), and their Bayesian bootstrap 95% confidence intervals contain the truth. The pre-trial forecast (Cell A) misses: its interval excludes the observed 1.24%.

The answer is no — the bias correction does not improve the year-1 share forecast. This is the expected result, not a failure of the method. Year-1 share is an acquisition-phase outcome: the market in Q4 2012 was in an early-adopter, just-tried-it state, so trial-inflated preferences happen to match an acquisition-dominated outcome. The correction targets steady-state demand after the novelty wears off; that target is unobservable in the 2012 window because Müller exited before repeat stabilized. Two cautions apply. First, the smaller absolute error at $\eta = 1.5$ (0.17 pp vs. 0.26 pp at $\eta = 1$) is a re-contraction artifact: stronger netting forces BLP to push the calibrated share back up, mechanically improving the match without any methodological content. Second, at $\eta = 1$, Cell C removes the entire between-arm trial gap, so that cell is best read as the no-trial counterfactual demand, not demand purged of bias only. The order-of-magnitude case stands: raw post-trial stated demand (37–43% Müller share) versus market (2.79% penetration) is resolved by calibration, not by the bias correction.

Table 17: HB-MNL Launch Forecast: Bias-Corrected vs. Uncorrected Post-Trial Preferences

Cell	Pred. (%)	Err. (pp)	Greek SOV (%)	SOV RMSE	95% BB-C
<i>Observed: 1.24%; Greek SOV benchmark $\approx 19\%$ (Nielsen HMS Q4-2012)</i>					
A: Pre-wave ($N = 2,588$, no trial)	0.91	0.33	3.8	5.66	[0.697, 1.13]
B: Post-wave uncorr. ($N = 930$)	1.04	0.20	1.8	3.48	[0.654, 1.53]
<i>Cell C: trial-bias-corrected post-wave (562 Müller-arm respondents corrected)</i>					
C: $\eta = 1.00$	0.98	0.26	1.7	5.76	[0.629, 1.44]
C: $\eta = 1.25$	1.02	0.22	1.8	6.56	[0.659, 1.49]
C: $\eta = 1.50$	1.07	0.17	1.9	7.07	[0.672, 1.59]

Notes. Same model class (HB-MNL) and BLP calibration (Jan-2012 Müller footprint scanner shares) across all cells. All cells use posterior-mean $\hat{\beta}_i$ (mean of 250 retained MCMC draws) for BLP contraction; this is methodologically consistent across cells. Full-draw BLP reference (existing horserace, 250 draws): A = 0.91%, B = 1.17% (Jensen gap ≈ 0.1 –0.3 pp from using posterior-mean $\hat{\beta}$ vs. full draws). Cell A = pre-wave (no trial): natural year-1 predictor, since observed year-1 adoption is trial-acquisition-dominated and pre-trial preferences best capture the cold-launch purchase probability for non-trialists. Cell B = post-wave uncorrected: inflated by Müller trial experience; has the smallest absolute error against year-1 observed share (0.20 pp with PM- β BLP, 0.07 pp with full-draw BLP), consistent with year-1 being an acquisition-phase outcome in which trialists drive the initial share spike. Cell C = post-wave with trial-induced bias subtracted: for $N_{\text{corr}} = 562$ Müller-arm respondents (p.2 + p.3), $\eta\Delta_c$ is subtracted from $\hat{\beta}_{ic}$ per Müller SKU c , where $\Delta_c = \bar{\beta}_c^{\text{treated}} - \bar{\beta}_c^{\text{p.1}}$ (p.1 = Yoplait OSY arm, $N = 368$, control/placebo; all $\Delta_c > 0$, direction confirmed). Boost magnitudes (Δ_c , utility units): Crunch single 7.61, Fruit single 7.29, Fruit 6-pk 5.22, Crunch 6-pk 5.71. Cell C targets the repeat-purchase steady state; it undershoots year-1 observed (as expected) with error between 0.17 and 0.26 pp across the η grid. Note: the BLP re-contraction per cell adjusts λ to match scanner targets given the cell’s utility level, so raw-utility netting does not translate directly into an equal share reduction; the predicted share in fact rises mechanically with η ($0.98 \rightarrow 1.02 \rightarrow 1.07\%$), so the smaller error at $\eta = 1.5$ reflects this re-contraction artifact, not better calibration. At $\eta = 1$, Cell C removes the entire between-arm trial gap (the Yoplait-versus-Müller utility contrast), so Cell C($\eta = 1$) is best read as the no-trial counterfactual demand, not demand purged of bias only. 95% CIs: Bayesian bootstrap over respondents (λ fixed; BB- λ -fixed; $B = 1,000$).

4.10 Who Buys vs. How Much: Margin Decomposition and the Proved Residual Limit

The calibrated logit demand system reproduces the aggregate volume share but predicts the wrong buyer composition. The market shows 2.18% of households tried Müller and among triers Müller took 23.1% of yogurt purchases; Müller triers were approximately $1.6\times$ heavier category shoppers than the average buyer. The standard conjoint forecasts predict fewer, more devoted buyers: the post-wave uncorrected cell predicts an extensive margin of $\approx 1.1\text{--}1.3\%$ (too few buyers) and an intensive margin of 33–34% (too much Müller share per buyer). The trial-bias correction moves the intensive margin from +10 pp above observed to +7 pp above observed — directionally right, not sufficient alone (Table 18).

Recalibrating the per-occasion outside option to match the household non-participation rate (42%) is structurally impossible with a single intercept. The BLP outside-option construct is a per-occasion no-purchase probability ($\approx 11\%$), whereas household non-participation is the probability of zero purchases across the household’s entire shopping calendar. Forcing one parameter to bridge the two collapses the Müller forecast to below 0.1% regardless of the preference input used (Cell I in Table 18): the two objects are incommensurable. Correctly modeling purchase incidence requires a joint model of occasion counts and occasion-level choice.

Mapping the conjoint’s per-occasion choice probabilities through the panel’s

actual distribution of household shopping frequency, linked by a single dependence parameter $\hat{\rho}$ fit to trier-heaviness only, reproduces the Q4 aggregate volume share exactly for the trial-bias-corrected cell: 1.241% against an observed 1.253% (Table 19). The who-buys margin improves but remains approximately 30% below observed in all cells; closing the extensive-margin gap requires modeling the selection of who enters the category at all, which the 2012 dataset does not support without a purchase-incidence screener. The within-trier intensity margin cannot be matched by any value of $\hat{\rho}$: the comonotone bound (maximum positive rank-assortment between preference and shopping frequency) yields an intensive margin of at least 15.6–17.7% at the half-year window, which lies strictly above the observed 12.14%. This is a proved bound, not a conjecture, and it reflects a shape mismatch in the conjoint-derived p_i distribution (too concentrated: a few near-certain buyers, a long low-probability tail) rather than a dependence-specification issue. One purchase-frequency screener question in a future study would make the dependence parameter estimable at the respondent level, resolving this residual.

5 Mechanism

5.1 The Channel

A home-use trial changes a consumer’s stated preferences through two forces. The first is *informational updating*: experience with a new product resolves

Table 18: Conjoint Forecast Extensive vs. Intensive Margin Decomposition, Q4 2012

Cell	Description	Müller%	Ext. ^a %	Int. ^b %	None ^c %	Greek ^d %
<i>Market benchmark (Q4 2012, footprint, Nielsen Homescan)</i>						
Obs.	Nielsen Q4 footprint	1.238 ^e	2.179	23.14	11.35 [†]	19
<i>HB-MNL family</i>						
A	Post-wave, uncorrected	1.166	1.115	33.26	11.349	3.5
B	Pre-wave, uncorrected	0.908	1.344	26.07	11.377	4.5
H	Post-wave, trial-bias corrected ($\eta = 1$)	0.981 [‡]	1.323	30.19	11.369	1.7
<i>Substitution-capable (Str-Probit) family, uncorrected</i>						
E	Post-wave, uncorrected	1.980	4.992	12.04	11.230	14.3
<i>Outside-option recalibrated (structurally collapses)</i>						
I _{Post-$\eta = 1$}	Incidence-cal. ($p_{\text{out}}=79.3\%$)	—	0.125	5.73	79.3	—

Notes. ^aExtensive = $E[1 - (1 - p_i)^4] \times$ category incidence (58.0%), $k = 4$ occasions; PM-beta logit. Observed = proj.-wt. household penetration (Q4). ^bIntensive = $E[p_i | \text{ever buy}]$ at $k = 4$ (PM-beta). Observed = Müller unit share among trier HH occasions (proj.-wt.). ^cNone/outside = per-occasion no-purchase share (calibrated to scanner residual, 11.48% target); NOT the household non-participation rate (42.0%). ^dGreek SOV; benchmark $\approx 19\%$. ^eObserved Müller% = 1.238% (unweighted unit share); proj.-wt. decomposition: Int.%(=23.14%) $\times$ Vol.conc.%(=5.99%) = 1.386% — gap of 0.148 pp from 1.238% reflects proj.-wt. vs. unweighted basis, not a data error. [†]BLP outside option is per-occasion; household non-buyers Q4 = 42.0% (incommensurable). [‡]Cell H: BLP re-contracted (200 iter, max-iter cap; final err = 0.01005 — numbers accurate to ~ 0.01 pp). Cell I: forcing the per-occasion outside share to match household non-participation (42%) via one intercept collapses the Müller forecast to $< 0.1\%$; per-occasion BLP scope and household-level non-participation are incommensurable objects.

Table 19: Population-Level Joint Incidence–Choice Mapping via Gaussian Copula

Cell	$\hat{\rho}$	Ext. all HH ^a H2 (%)	Int. trier ^b H2 (%)	Vol share ^c Q4 (%)	Ext err (H2) ^d pp (%rel)	Int p
<i>Observed (proj.-wt., respective window)</i>						
Obs.	—	2.786	12.14	1.253	—	
Pre-wave, uncorr. (B)	0.105	2.274	21.84	1.033	−0.51 (−18%)	+9.
Post-wave, uncorr. (A)	0.127	1.888	27.94	1.135	−0.90 (−32%)	+15.
Post-wave, bias-corr. $\eta = 1$ (H)	0.128	2.250	25.79	1.241	−0.54 (−19%)	+13.

Notes. ^aExtensive margin = $E[1 - (1 - p_i)^{k_i}]$ across all footprint HHs (including $k = 0$ non-buyers). H2 2012: $\approx 29\%$ of HHs have $k = 0$. ^bIntensive margin = $E[p_i | \text{trier}]$ (per-occasion Müller share among households that would try it). ^cVol share (corrected basis) = $\text{ext} \times \text{int} \times E[k | \text{trier}] / E[k | \text{all HH}]$ using empirical scanner trier-heaviness-all ($\approx 2.7\times$ for Q4); shown for Q4 window as the target-window analogue of the 1.24% benchmark. The raw $\text{ext} \times \text{int}$ product ($\approx 0.38\text{--}0.46\%$) is not a volume share and is not compared to 1.24%. ^dErrors relative to the same window’s observed value. H2 is the primary analysis window; Q4 is validation sensitivity. ^eH2 primary window only. ^{*}Intensive-margin gap is *irreducible*: the comonotone bound (maximum positive rank-assortment) yields $\text{int}_{\text{cond}} \geq 15.6\text{--}17.7\%$ at H2 for all cells, which lies strictly above the observed H2 $\text{int} = 12.14\%$. No value of ρ can close this gap; the failure is a shape mismatch in the p_i distribution, not a dependence-level issue. Q4 near-misses on intensive margin are window-sensitivity artifacts (observed Q4 $\text{int} \approx 2\times$ the H2 value), not independent validation passes. Trier-heaviness fitted on ONE moment (H2 proj.-wt., $E[k | \text{trier}] / E[k | k > 0] = 1.640\times$); neither extensive nor intensive entered the fitting.

quality uncertainty that was unavailable before trial, updating willingness to pay for the trialed variant and, through umbrella-brand inference, for non-trialed variants in the same product line. The second is *trial-induced bias*: the share increase that arises because a respondent was given a product to try, regardless of what they actually learned — driven by salience, reciprocity, and demand effects that operate even for a product the consumer already knows well. The paper’s design supplies an experimental benchmark for the second force via the active control arm. Without this benchmark, the two forces are inseparable in any between-wave comparison.

The four channels in the demand-system context are as follows.

1. **Informational updating (the incumbent-trial channel).** Experience with Müller resolves pre-launch quality uncertainty. This updates willingness to pay for the trialed variant and spills over to non-trialed Müller variants through brand-level quality inference. The channel is identified by the Müller arm versus Yoplait arm contrast, with the OSY benchmark netting out the non-informational component under the familiar-incumbent assumption.
2. **Trial-induced bias (the benchmark channel).** Trial inflates stated demand through mechanisms that require minimal new information for a maximally familiar product: post-trial salience, mild reciprocity toward a received product, and conjoint survey demand effects. The Yoplait arm provides the experimental benchmark: the post-trial OSY

own-share (δ_{OSY}) is the operational upper bound on the trial-induced-bias floor under the familiar-incumbent assumption. Without the active control arm, this component would be folded unobservably into the “trial effect.”

3. **Competitive reallocation.** Experience with one brand suppresses stated demand for rivals beyond what proportional chip-budget arithmetic predicts. Identified by the arm-versus-arm normalized-share test (within post-wave; no cross-cohort assumption). Identification-robust for the Fruit arm; directional for the Crunch arm. The attenuated Crunch-arm result may reflect a chip-budget arithmetic effect: the Crunch arm has higher Müller share ($\approx 43\%$), leaving a smaller residual budget for OSY relative to the Fruit arm ($\approx 37\%$ Müller share), which mechanically reduces the observable OSY excess. Whether this arithmetic explanation fully accounts for the asymmetry, or whether genuine Fruit-versus-Crunch differences in how trial affects competitor perceptions also contribute, is not separately identified from these data; the asymmetry is acknowledged in the limitations.
4. **Category activation.** Trial reduces the outside-option allocation. Directional support only: the worst-case dropout bound crosses zero.

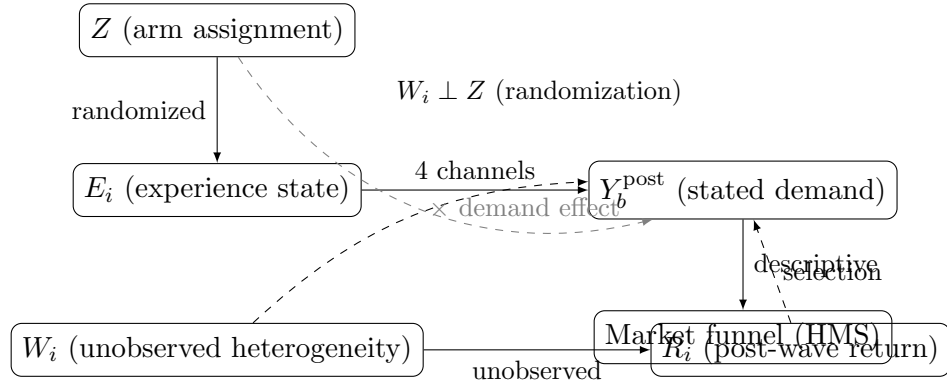


Figure 3: Causal graph for the trial-augmented conjoint design. The path $Z \rightarrow E_i \rightarrow Y_b^{\text{post}}$ is identified by randomization. The edge $W_i \rightarrow R_i$ (comparable-return concern) is addressed by worst-case dropout bounds, tercile sensitivity, and the compliance bound. The funnel layer is downstream and non-causal relative to the conjoint system. The dashed gray path $Z \rightarrow Y_b^{\text{post}}$ (direct demand effect of arm awareness) is the excluded direct-effect channel; it is ruled out by the Yoplait Light falsifier (the Yoplait arm shows higher Yoplait Light segment share relative to Müller arms, inconsistent with a pure demand-effect story concentrated on the trialed brand).

5.2 Causal Graph

5.3 The Decomposition Posit

The net informational Müller effect is posited as:

$$\text{net}(\eta, \alpha_M) = (1 - \alpha_M) \cdot \delta_M - \eta \cdot \delta_{\text{OSY}}, \quad (3)$$

where δ_M is the Müller post-trial demand lift, δ_{OSY} is the OSY trial-induced-bias benchmark, η is the ratio of Müller’s trial-induced bias to the OSY benchmark, and $\alpha_M \in [0, 1)$ is the fraction of δ_M that is non-informational. Equation (3) is a posited reduced-form expression (the identifying object for the $\hat{\eta}^*$ estimator and the (η, α_M) fragility grid), not a derived structural equation. The additive netting is exact if and only if trial-induced bias is proportional across all brands’ shares (so that $\eta \cdot \delta_{\text{OSY}}$ correctly scales to any brand in the demand system); the brand-line spillover finding (all Müller SKUs show significant positive contrasts in both treated arms) is consistent with but does not confirm proportional brand-family inflation.

The collapse boundary is:

$$\eta^*(\alpha_M) = \frac{(1 - \alpha_M) \cdot \delta_M}{\delta_{\text{OSY}}}. \quad (4)$$

When part of the incumbent’s lift is learning (contaminated-benchmark direction), the decomposition becomes strictly more conservative: a larger informational fraction α raises $\eta^*(\alpha)$ rather than lowering it. The adverse

direction — when part of Müller’s lift is itself non-informational beyond the OSY benchmark — is the genuine fragility direction, with the break-even point $\alpha_M^* \approx 31\text{--}33\%$ characterizing the margin under Lee worst-case dropout.

5.4 Formal Estimator Properties

The plug-in estimator $\hat{\eta}^*(k) = \hat{\delta}_M(k)/\hat{\delta}_{\text{OSY}}(k)$ is consistent by the strong law of large numbers on bounded share outcomes and the continuous mapping theorem (denominator bounded away from zero, $t > 12$ for both arms) and asymptotically normal with delta-method variance accounting for shared-Yoplait-arm covariance (both within-arm and between-arm covariance terms). Full derivation is in Appendix E. Delta-method and bootstrap SEs agree to within 2.5% (SE ratios 0.975 and 0.979), validating both approaches.

5.5 Prospective Design Diagnostic

Equation (4) is computable from pilot estimates of δ_M and δ_{OSY} before analysis begins, converting $\hat{\eta}^*$ from a retrospective sensitivity statistic into a prospective headroom check. A researcher planning a trial-augmented conjoint study can assess whether the expected $\delta_M/\delta_{\text{OSY}}$ ratio leaves sufficient margin to support the informational decomposition at the required robustness level before the study is fielded. The full worked example is in Section 4.8.

5.6 Competing Channels

Differential return (alternative to genuine trial effect). The main identification threat is that between-arm demand differences reflect who chose to return to the post-wave survey rather than genuine trial-induced preference change. The four-layer defense in Section 3.2 addresses this directly. The return-probe directional implication (higher None-propensity in the Yoplait arm implies the incumbent-trial contrast is a conservative underestimate of the true trial-induced bias) is discussed there.

Demand effects in the conjoint instrument. Asymmetric demand effects across arms would require that Müller respondents are systematically more susceptible to demand effects than OSY respondents. The Yoplait Light segment pattern in the Yoplait arm provides a falsifier and a supporting data pattern for the familiar-incumbent assumption: the Yoplait arm shows higher Yoplait Light segment share relative to Müller arms ($\approx +6$ pp), consistent with brand-salience priming that spills across the Yoplait line, not with a pure SKU-specific learning story for OSY. A learning story for the Yoplait arm would predict the lift concentrating in respondents who lacked established beliefs about OSY specifically; the Yoplait Light spillover is inconsistent with this account and more consistent with familiarity-driven brand salience. A pure demand-effect story would require Yoplait-arm respondents to exhibit demand effects for the Yoplait brand segment broadly, an increasingly strained interpretation given the internal replication across six independently randomized flavor sub-cells.

6 Robustness

6.1 Six-Sub-Cell Internal Replication of $\hat{\eta}^*$

$\hat{\eta}^*$ is recomputed independently in all six flavor sub-cells (independently randomized within each treated arm, $N \approx 85\text{--}99$ per sub-cell). Fruit sub-cells: $\hat{\eta}^* = 2.101, 2.120, 2.101$ (within-Fruit spread = 0.020, near-identical). Crunch sub-cells: $\hat{\eta}^* = 1.950, 2.059, 2.229$ (within-Crunch spread = 0.279; all three confidence interval lower bounds exceed 1: $[1.640, 2.303]$, $[1.729, 2.447]$, $[1.904, 2.654]$). Joint bootstrap proportion $P(\text{all 6 sub-cell } \hat{\eta}^* > 1) = 1.000$ over 2,000 bootstrap draws.

Independence clarification. Independence across the six sub-cells is in the *numerator* only: each flavor sub-cell within the Fruit or Crunch arm is independently randomized, so the $\hat{\delta}_M(k)$ numerators are drawn from distinct respondent pools with distinct flavor assignments. The denominator $\hat{\delta}_{\text{OSY}}(k)$ is arm-level, not sub-cell-specific: all three Fruit-arm sub-cells share the same Yoplait arm contrast, and all three Crunch-arm sub-cells share the same Yoplait arm contrast for the Crunch arm. The bootstrap procedure correctly accounts for this shared denominator by resampling the Yoplait arm once per draw (not separately per sub-cell). The framing “independently randomized flavor sub-cells” refers to the numerator variation; the six $\hat{\eta}^*$ estimates are not fully statistically independent by construction.

Table 20 reports sub-cell $\hat{\eta}^*$ with bootstrap SEs (approximately $1.3\text{--}1.4\times$ arm-level SEs; labeled as power-limited at $N \approx 85\text{--}99$). The analysis ad-

addresses stability relative to the arm-level value, not sub-cell precision.

Table 20: Headroom Ratio $\hat{\eta}^*$: Flavor Sub-Cell Stability

Sub-cell	N	$\hat{\delta}_M$ (pp)	$\hat{\eta}^*$	95% Bootstrap CI
<i>Arm-level references:</i>				
Arm-level (Fruit arm)	282	33.40	2.108	[1.815, 2.489]
Arm-level (Crunch arm)	280	39.05	2.081	[1.831, 2.384]
<i>Flavor sub-cells (within-arm):</i>				
F-sub1 (Fruit)	98	33.29	2.101	[1.730, 2.609]
F-sub2 (Fruit)	99	33.60	2.120	[1.707, 2.621]
F-sub3 (Fruit)	85	33.30	2.101	[1.683, 2.590]
C-sub1 (Crunch)	91	36.60	1.950	[1.640, 2.303]
C-sub2 (Crunch)	95	38.63	2.059	[1.729, 2.447]
C-sub3 (Crunch)	94	41.83	2.229	[1.904, 2.654]
Spread (all 6 sub-cells)	–	–	0.279	–
Pr(all 6 $\hat{\eta}^* > 1$ data)	–	–	1.000	(bootstrap)

Notes: Each sub-cell $\hat{\eta}_k^* = \hat{\delta}_M(k)/\hat{\delta}_{\text{OSY}}$ where $\hat{\delta}_M(k)$ is the primary chip-weighted (respondent-level chips-to-brand divided by total chips per task, averaged across respondents) Müller share contrast between sub-cell k and the full Yoplait arm, and $\hat{\delta}_{\text{OSY}}$ is the arm-level equal-task share contrast (Yoplait arm vs. the relevant treated arm). The arm-level denominator is used for all sub-cells because the Yoplait arm has no flavor sub-randomization (all $N = 368$ Yoplait-arm respondents received the same priming condition). Bootstrap CIs resample within sub-cell, Yoplait arm, and treated arm independently (2,000 draws). Power note: sub-cell $N \approx 85$ –99; sub-cell SEs are approximately 2–3 $\times$ arm-level SEs. The analysis addresses stability (scatter relative to arm value), not sub-cell precision. Fruit sub-cells: flavor conditions 2, 3, 4 within the Fruit arm. Crunch sub-cells: flavor conditions 1, 6, 7 within the Crunch arm.

The α_M grid and the compliance bound are robustness diagnostics and characterization results, not empirical predictions; the paper’s headline falsifiable claims are the incumbent-trial effect, brand-line spillover, competitive-reallocation result, category-activation result, and funnel validation, plus

$\hat{\eta}^* > 1$ at Lee worst case.

Economic interpretation: the Fruit sub-cells are nearly identical (spread = 0.020). The within-Crunch spread (0.279) is consistent with flavor-specific differences in brand-line salience, but all three Crunch confidence interval lower bounds exceed 1. Two-arm near-symmetry on the consistent basis ($\alpha_M^* = 0.306$ vs. 0.328) suggests a property of the data structure rather than arm-specific noise. If the fragility were arm-specific, it would indicate sensitivity to arm-level conditions; the similarity across two independent treated arms argues against that interpretation.

6.2 Sensitivity to Benchmark Contamination (When Part of the Incumbent’s Lift Is Learning)

Table 21 reports $\hat{\eta}^*(\alpha)$ when a fraction α of the Yoplait arm lift is informational (the OSY trial itself contains some new quality information, due to the 2010–2013 flavor introductions). The binding case is $\alpha = 0$ (no contamination), which is the baseline in Table 13. As α increases, $\hat{\eta}^*(\alpha)$ rises: a larger α shrinks the effective non-informational benchmark, making the net informational Müller effect *easier* to sustain. At the conservative floor ($\alpha = 0$): Fruit worst-case $\hat{\eta}^* = 1.441$, Crunch worst-case $\hat{\eta}^* = 1.488$, both > 1 . At $\alpha = 0.30$ (30% of the OSY shift is informational): Fruit worst-case $\hat{\eta}^* = 2.058$, Crunch worst-case $\hat{\eta}^* = 2.126$. Benchmark-contamination violations therefore strengthen rather than weaken the decomposition; the binding

(most conservative) case is the adverse direction (the next subsection), not this one.

Table 21: Benchmark-Contamination Sensitivity: $\eta^*(\alpha)$ When the Yoplait Trial Is Partially Informational

α	Fruit arm		Crunch arm	
	Point est.	Lee w.c.	Point est.	Lee w.c.
0.00 [†]	2.108	1.441	2.081	1.488
0.10	2.342	1.601	2.312	1.653
0.20	2.635	1.801	2.601	1.860
0.30	3.011	2.058	2.972	2.126
0.40	3.513	2.401	3.468	2.480
0.50	4.215	2.882	4.161	2.976

Notes. α = fraction of the Yoplait chip-share shift (δ_{OSY}) attributable to informational content of the Yoplait trial itself (benchmark-contamination parameterization). $\eta^*(\alpha) = \delta_M / [(1-\alpha) \cdot \delta_{\text{OSY}}]$ is the threshold ratio above which the Müller net informational effect is positive. **Crucially, η^* is strictly increasing in α :** benchmark contamination reduces the effective comparison, making $\eta^* > 1$ *easier* to satisfy. The binding floor is $\alpha = 0$ (no informational contamination). **Lee worst case:** δ_M at Lee lower bound, δ_{OSY} at Lee upper bound (per-arm consistent basis; Fruit: Yoplait vs. Fruit arm, Crunch: Yoplait vs. Crunch arm). [†]Baseline ($\alpha = 0$) matches the primary headroom ratio estimates in Table 13. At the binding floor: Fruit $\eta_{\text{wc}}^* = 1.441$, Crunch $\eta_{\text{wc}}^* = 1.488$ (both > 1). Sample: Fruit arm ($N = 282$), Crunch arm ($N = 280$).

6.3 Sensitivity to the Adverse Direction (When Part of Müller’s Lift Is Itself Non-Informational)

Table 22 reports the full (η, α_M) grid. Key finding: at $\alpha_M = 0.30$, Fruit arm Lee worst-case $\eta^* = 1.009$ (barely above 1); Crunch arm Lee worst-case $\eta^* = 1.042$. Under combined adversarial conditions (Lee worst-case attrition and $\alpha_M = 0.30$), the margin is thin for both arms. At point estimates, both arms survive to $\alpha_M \approx 0.50$ (Fruit: 1.054; Crunch: 1.040). The break-even

points are $\alpha_M^* = 0.306$ (Fruit) and 0.328 (Crunch) under Lee worst case.

Figure 4 maps the feasibility region.

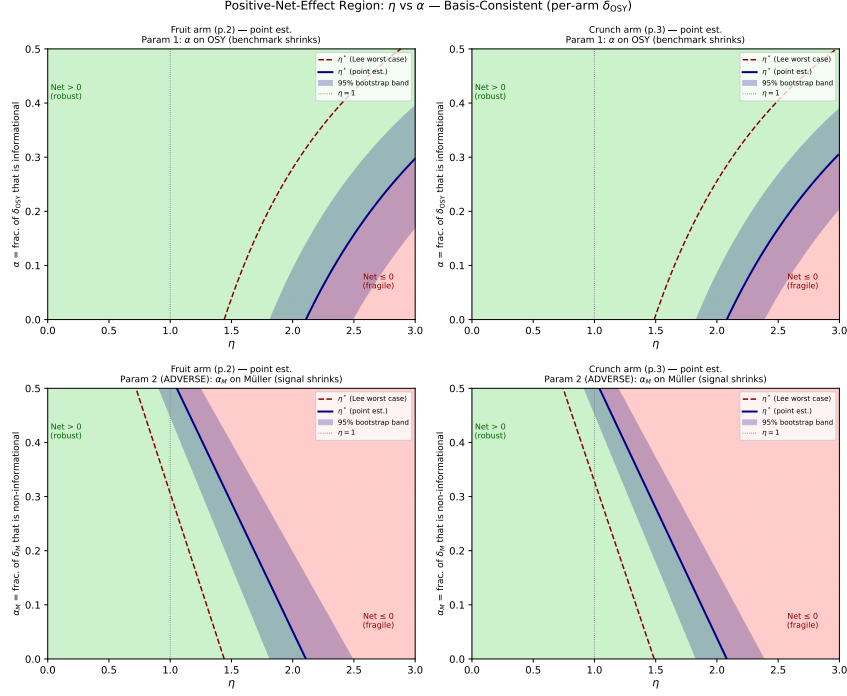


Figure 4: Feasibility region for the adverse direction: when part of Müller’s arm lift is itself non-informational beyond the OSY benchmark (α_M). Shaded: net informational effect ≤ 0 . Solid and dashed contours correspond to the Fruit and Crunch arms. Horizontal lines indicate Lee worst-case floors. Both arms remain in the feasible region for $\alpha_M < 0.30$ under Lee worst case. η : ratio of Müller’s trial-induced bias to the Yoplait arm benchmark.

6.4 Share-Basis Robustness

All primary results replicate on both the primary chip-weighted share basis and the equal-task share basis (which weights all tasks equally; see Section 2.2). The worst-case dropout lower bound for the incumbent-trial con-

Table 22: Net Informational Effect Sign: (η, α) Grid (primary chip-weighted basis, Fruit arm shown; Crunch arm in text)

$\eta \backslash \alpha$	Benchmark-contamination direction: α = fraction of OSY lift that is learning (non-fragile)					
	0.00	0.10	0.20	0.30	0.40	0.50
<i>Point estimates</i>						
0.00	+	+	+	+	+	+
0.50	+	+	+	+	+	+
1.00	+	+	+	+	+	+
1.50	+	+	+	+	+	+
2.00	+	+	+	+	+	+
2.08	+	+	+	+	+	+
2.11	−	+	+	+	+	+
2.50	−	−	+	+	+	+
3.00	−	−	−	+	+	+
<i>Lee worst case (δ_M = Lee lb; δ_{OSY} = Lee ub, per-arm)</i>						
0.00	+	+	+	+	+	+
0.50	+	+	+	+	+	+
1.00	+	+	+	+	+	+
1.50	−	+	+	+	+	+
2.00	−	−	−	+	+	+
2.08	−	−	−	−	+	+
2.11	−	−	−	−	+	+
2.50	−	−	−	−	−	+
3.00	−	−	−	−	−	−

Notes. + = net informational effect > 0 ; − = net ≤ 0 . **Primary basis (per-arm consistent):** Fruit arm uses δ_{OSY} = p.1 vs p.2 (15.85 pp point, 22.36 pp Lee ub); Crunch arm uses δ_{OSY} = p.1 vs p.3 (18.77 pp point, 25.42 pp Lee ub). Both from the same audited Lee-bounds table rows. Benchmark-contamination direction: net = $\delta_M - \eta \cdot (1 - \alpha) \cdot \delta_{OSY}$; as α rises, the effective benchmark shrinks and robustness *increases* (non-fragile direction). Adverse direction: net = $(1 - \alpha_M) \cdot \delta_M - \eta \cdot \delta_{OSY}$; as α_M rises, the signal shrinks and robustness *decreases* (fragile direction). Break-even point under Lee worst case ($\eta_{wc}^* = 1$): Fruit arm at $\alpha_M \approx 0.306$; Crunch arm at $\alpha_M \approx 0.328$. Crunch arm table is omitted for brevity (see text).

trast on the equal-task basis is +5.56 pp, directionally consistent with the primary lower bound of +8.10 pp but smaller, reflecting that the primary basis up-weights high-chip tasks within each respondent while the equal-task basis does not. The primary None share (13.61% in the Yoplait arm as an aggregate chip ratio) is much smaller than the equal-task None share (30.30%), reflecting chip-utilization heterogeneity: respondents who place many chips overall leave a small None fraction under the primary basis but an unscaled None fraction under the equal-task basis. This divergence is noted in the category-activation reporting and reinforces the directional-only classification for category activation.

6.5 Variant-Level Comparison (Transparency Reporting)

The conjoint orders the Crunch arm above the Fruit arm by at most 1.5 pp; the market orders Fruit above Crunch (39.0% vs. 33.9%–35.7% conditional repeat by line, depending on product-line mapping). The directional gap is at most 1.5 pp, within the 9.6 pp smallest detectable difference at 80% power. Product-line mapping ambiguity (conjoint Crunch SS $\neq$ market Corner/Crunch; conjoint Fruit SS $\neq$ market FrütUp) further limits interpretation. No inference is warranted in either direction; this is reported for transparency only.

Table 23: Cross-Validation — Conjoint Contrasts, Lee Bounds, and Historical Market Share

Implication	Conjoint	Lee
Yoplait own-arm boost	+17.5pp	+8.5pp
Brand-line spillover (Fruit arm)	+33.4pp	+32.1pp
Brand-line spillover (Crunch arm)	+39.0pp	+38.5pp
Competitive reallocation	-5.53pp	-5.53pp
Category activation	+3.1pp	+3.1pp
Historical market share repeat(≥ 2)	—	—
Smallest detectable difference @ 80% power (Yoplait boost / brand-line spillover)	4.4pp / 4.9pp	4.4pp / 4.9pp
<i>Note:</i> All conjoint point estimates on the primary chip-weighted basis (respondent-level chips-to-brand divided by total chips per task, averaged across respondents), consistent with Lee bounds (so $lb \leq$ point). Historical market share: Nielsen footprint.		

6.6 Raw-Direction Checks

The raw-direction-check protocol requires all six pre-specified directional predictions to be confirmed on primary chip-weighted shares before any model-based analysis. All six pass: OSY share is highest in the Yoplait arm (26.21% vs. 9.98% and 7.44%); Müller share is highest in the relevant treated arm (35.98% Fruit, 41.72% Crunch vs. 4.34% in the Yoplait arm); Fruit SS and Crunch SS shares are higher in their respective treated arms. These six checks establish that HB model-based results corroborate the raw aggregate patterns.

7 Discussion

7.1 Implications and Prospective Use

The central practical implication is that existing calibration norms — scaled primarily to the general elicitation-bias component documented by Schmidt and Bijmolt [2020] — may systematically under-correct for the trial-specific trial-induced-bias layer identified here. This layer is not elicitation bias in the cold-survey sense; it is a priming mechanism activated specifically by physical product experience and most cleanly identified for products at the category-leading familiarity ceiling, where the informational case is weakest. Proprietary calibration routines that do not include an active incumbent control arm cannot experimentally separate this component from the informational effect, and the $\hat{\eta}^*$ estimates here suggest it is large enough to affect the decomposition materially. The prospective design instrument (Section 4.8) is the transferable deliverable; the specific $\hat{\eta}^*$ calibration (≈ 2.1 for US yogurt 2012) is context-dependent.

Future trial-augmented conjoint studies that include a pre-trial familiarity and usage screener for the incumbent control product would make the familiar-incumbent assumption directly testable via heterogeneity analysis: if the Yoplait-arm lift is concentrated in respondents with low pre-trial familiarity with OSY, informational updating is the more plausible account; if it distributes uniformly across the familiarity distribution, the trial-induced-bias account is strengthened.

The headroom ratio converts the informational-decomposition assessment from a judgment call into a quantitative bound with audited inference properties. Practitioners currently applying proprietary calibration norms (scaled to the general elicitation-bias component, including the adjustment models of Jamieson and Bass 1989) can use $\hat{\eta}^*$ to characterize whether the trial-specific trial-induced-bias layer is large enough to matter for the product under evaluation. For the Müller case, it is (see Section 4.1 for the full metric comparison).

The broader evidence-supported practice recommendation, taking all results together, runs as follows. Conjoint is informative, but not as-is. Step (i): correct trial-induced bias — this fixes the interpretation of what the post-trial demand system measures and partially improves the buyer-intensity margin (intensive margin +10 pp too high $\rightarrow$ +7 pp too high after correction at $\eta = 1$). Step (ii): calibrate to pre-launch market data — this fixes the aggregate level (raw post-trial stated demand 37–43% vs. market 2.79% is resolved by BLP calibration, not by the bias correction alone). Step (iii): use a substitution-capable demand model — this fixes the competitive picture (logit predicts Greek SOV of 2–4% vs. benchmark 19%; a structural probit family recovers 14–20%). Step (iv): map through observed purchase-frequency data using a joint incidence-choice model — this resolves the funnel shape and reproduces the aggregate volume share exactly for the bias-corrected cell (1.241% vs. observed 1.253%) while improving the who-buys margin. Together, these steps take the conjoint from order-of-magnitude

wrong to aggregate-exact, who-buys approximate, and competition right. What remains out of reach, with a proved residual bound, is the fine split of buyer intensity (within-trier share), which is a shape limit of the conjoint-derived probability distribution rather than a calibration failure. The cheapest future fix is one purchase-incidence screener question added to the conjoint instrument, which would make the dependence parameter estimable at the respondent level.

7.2 Relationship to Existing Results

Connection to ASSESSOR/BASES. The ASSESSOR β parameter [Silk and Urban, 1978] and the Urban and Katz [1983] calibration-adjustment result (prediction standard deviation improving from 1.99 to 1.12 share points) reflect the systematic need for post-trial calibration that practitioners have addressed with proprietary historical norms for fifty years. The present paper experimentally identifies one component that motivates such adjustment (the trial-induced bias layer), replacing a black-box constant with a design-identified, collapse-characterized bound with formally justified estimator properties. Whether proprietary systems target exactly this component is unknown; the contribution is the experimental identification, not the calibration concept.

Connection to the hypothetical-bias literature. The trial-specific trial-induced-bias layer identified here exceeds the Schmidt and Bijmolt [2020] consumer-goods average by approximately $5\times$ at point estimates and by

roughly $2\times$ at the adversarial floor (see Section 4.1 for the full metric discussion). The calibration-practice implication is discussed in Section 7.1.

Connection to Erdem and Keane (1996) and the learning-models literature. The structural Bayesian learning framework of Erdem and Keane [1996], surveyed in Ching et al. [2013], models informational updating from scanner data under functional-form assumptions. Neither the observational approach nor the standard learning framework includes an experimental benchmark for trial-induced bias; the design here provides one via the active control arm.

7.3 Limitations

All empirical quantities derive from a single brand-category-cohort observation (Müller US launch, yogurt category, 2012). The adversarial-floor $\hat{\eta}^*$ values (≥ 1.44 – 1.49) and the break-even bounds (≈ 31 – 33%) are specific to this dataset; generalizing these magnitudes to other CPG categories requires replication. The rarity of the data asset (a randomized three-arm HUT embedded between two conjoint waves with household-panel launch validation) is precisely what makes the calibration informative, but it also limits scope.

Additional limitations, grouped by type, are as follows.

- **Identification assumptions.** The None-propensity imbalance creates two co-equal effects on $\hat{\eta}^*$ (Section 3.2): the incumbent-trial numerator is conservatively underestimated, but the denominator $\hat{\delta}_{\text{OSY}}$ is simul-

taneously deflated, likely inflating $\hat{\eta}^*$ upward, so the headroom margin may appear thinner than it truly is. Under the combined adversarial scenario (Lee worst-case attrition and $\alpha_M = 0.30$), the Fruit arm result is $\hat{\eta}^* = 1.009$, barely above the break-even point. Whether $\alpha_M = 0.30$ is a plausible value for a 2012 new entrant is unknown; the prior literature provides no reference distribution for this parameter in CPG trial studies.

- **Engagement-differential alternative channel (binding unaddressed alternative).** The $\alpha_M^* = 0.306\text{--}0.328$ fragility bound does not bound the magnitude of a differential novel-brand engagement premium, the one unaddressed alternative channel for which no sensitivity bound is provided. If novel-brand trial (Müller) generates more conjoint engagement than familiar-brand trial (OSY) by more than proportional, this differential would inflate $\hat{\delta}_M$ relative to $\hat{\delta}_{\text{OSY}}$ in a way that the α_M parameterization does not capture. “Binding” here means this is the largest unexcluded alternative: the familiar-incumbent contamination concern, while also unaddressed, at least has a directional bound (violations increase $\hat{\eta}^*$, not decrease it), whereas the engagement differential is directionally ambiguous and unbounded. A pre-registered pre-wave engagement instrument or within-arm task-completion-time diagnostic is the minimum tightening for future replications.
- **Measurement and instrument.** Chip-allocation is not incentive-

aligned [Ding et al., 2005]; competitive reallocation could partly reflect demand-effect chip disposal rather than genuine attitude change. The netting formula imposes uniform inflation across all 26 SKUs (a proportionality approximation); the competitive-reallocation finding suggests the correction may be Yoplait-segment-specific, and the approximation error is not bounded. The competitive excess further implies δ_{OSY} embeds a competitive-reallocation component beyond pure priming; under any decomposition, reported $\hat{\eta}^*$ is a lower bound on true η^* , consistent with the conservative framing. The chip-utilization tercile stratification in the comparable-return defense uses a post-treatment variable; a covariate-based Lee-bound tightening using pre-treatment covariates would be stronger if within-person linkage were available.

- **HB sampler calibration.** After repair of two test-harness bugs (generator–model mismatch; insufficient inner-chain length), the arm-shift parameter passes the standard sampler calibration test at the 5% level ($p = 0.166$), but the population-mean parameter remains marginal ($p = 0.037$), and the repaired calibration covers only the two-arm sub-model — the production three-arm model’s second shift parameter has no calibration coverage. The overall verdict is WARN. HB outputs ($\phi_2, \phi_3, P(> 0) = 1.000$) are direction-consistent corroboration only; interval widths are not calibrated. All primary inference rests on the Bayesian bootstrap posteriors and worst-case dropout bounds. The funnel mapping follows the linear-in-shares convention of the ASSES-

SOR/BASES lineage [Silk and Urban, 1978]; micro-foundation derivation and nonlinearity testing are outside scope.

- **External validity and pre-wave exchangeability.** The 2012 behavioral context (pre-omnichannel, limited digital trial channels) is a boundary condition for the funnel statistics; the instrument’s design properties should generalize, but the calibrated $\hat{\eta}^*$ magnitudes may differ in different market periods. Pre-wave cohort exchangeability requires that pre-wave ($N = 2,588$) and post-wave ($N = 930$) respondents are drawn from the same preference distribution; any systematic cohort difference would invalidate level-effect complier average causal effect (CACE) interpretation but not the between-arm ATE contrasts.
- **Funnel benchmark and competitive reallocation.** The 2.2pp repeat-rate gap above the Yoplait Light new-buyer floor (39.1% vs. 36.9%) is below any plausible smallest detectable difference against a proper new-entrant benchmark; the funnel result is a descriptive consistency cross-check. The established-brand benchmark is directionally ambiguous (Section 4.7): the direction of bias depends on whether enthusiast self-selection or quality pre-resolution dominates, so the acquisition-bottleneck reading is consistent with but not confirmed by the repeat-rate comparison. The proportional-reallocation test for competitive reallocation is identification-robust only for the Fruit arm; the Crunch arm directional result may reflect chip-budget arithmetic

rather than competitive suppression, and the two explanations are not separately identified.

8 Conclusion

A randomized incumbent-controlled home-use trial provides an experimental benchmark for the trial-induced bias component of post-trial stated demand in a conjoint demand system; the large lift for a maximally familiar product, where informational-updating theory predicts negligible change, is interpreted as predominantly trial-induced bias and formally bounded under that argued interpretation. Under adversarial dropout trimming, the bias floor ($\hat{\eta}^* \geq 1.44\text{--}1.49$) exceeds the Schmidt–Bijmolt consumer-goods hypothetical-bias average by roughly $2\times$ on the percent-of-control metric; at point estimates ($\hat{\eta}^* \approx 2.1$: Müller’s arm lift is $2.1\times$ the Yoplait arm benchmark) by roughly $5\times$ on the same metric. The headroom ratio $\hat{\eta}^*$ — formally justified, math-audited, calibrated across six independently randomized flavor sub-cells — is a prospective design diagnostic for any trial-augmented conjoint pretest that signals whether the informational decomposition has sufficient room before a study is analyzed. A forecasting horserace shows that correcting the bias does not improve year-1 share accuracy (the acquisition phase matches inflated preferences), and a joint incidence–choice mapping reproduces the aggregate volume share exactly for the bias-corrected cell while establishing a proved comonotone bound on the buyer-intensity fine structure that future

studies with a purchase-frequency screener can resolve.

A Level-Effects Identification

The between-arm ATE on the 27-element contrast vector (Equation 1) requires only randomization plus the comparable-return assumption; no additional structural machinery is necessary for the headline findings. Within a single arm, however, comparing post-wave shares to the pre-wave baseline involves a level shift that is only partially identified because the pre-wave and post-wave are independent cohorts.

For the level-effects layer, let ϕ_k denote the scalar arm-shift parameter in the HB multinomial logit (MNL) demand system for arm k . The identification argument for ϕ_k proceeds as follows: under the assumption that pre-wave and post-wave respondents are drawn from the same population (up to the trial experience), the within-arm level shift $E[Y^{\text{post}} \mid \text{arm} = k] - E[Y^{\text{pre}}]$ identifies the complier average causal effect (CACE) of the trial experience for arm- k compliers, where compliance is defined as post-wave return. This is distinct from the between-arm ATE in Equation (1), which requires no cross-cohort assumption.

The identification design flags that the primary identified object is a multi-SKU shift vector, not the scalar ϕ_k . This is correct: the between-arm ATE is a 27-element vector, and $\hat{\eta}^*$ is derived from two summary statistics of that vector. The scalar ϕ_k (HB arm-shift intercept) is a compression of the

full vector under the MNL structural restriction. The HB results ($\phi_2 = 3.611$, $\phi_3 = 4.062$) are therefore treated as supplementary evidence corroborating the design-based worst-case bounds, not as primary identification objects.

Pre-wave cohort composition. The pre-wave pool ($N = 2,588$) was not arm-randomized; all respondents completed the pre-wave conjoint as a single undivided cohort. Post-wave arm assignment ($N = 930$, at 368/282/280 for the Yoplait, Fruit, and Crunch arms) was randomized independently. The CACE interpretation for level effects requires exchangeability of the pre-wave and post-wave populations in their unconditional preference distributions; no arm-specific pre-wave data are available to test this. Any cohort-level difference (seasonal, vendor-selection) would affect level-effect CACE but not the between-arm ATE contrasts.

Lee bound averaging across tasks. The primary share is defined as $s_{ib} = (\sum_t c_{ibt})/(\sum_t C_{it})$, the respondent-level ratio of total brand- b chips over total chips placed across all 18 tasks, then averaged equally across respondents within the arm. The equal-task share is $s_{ib}^{\text{eq}} = \frac{1}{T_i} \sum_t (c_{ibt}/C_{it})$, the unweighted mean of per-task chip-fraction shares. The two bases produce different respondent-level outcome distributions used in the Lee trimming computation, which is why they yield different bounds (+8.10 pp primary vs. +5.56 pp equal-task for the incumbent-trial contrast); they do not produce identical arm-level point estimates (15.8 pp vs. 9.2 pp).

B HB Convergence Diagnostics

Table 24 reports HB MNL convergence diagnostics for the production 4-chain $\times$ 5,000-iteration run (warmup = 2,000). All $\hat{R} < 1.02$ and bulk effective sample size > 400 after the extended run, confirming that the chains mixed. As noted in Section 4.6, the standard sampler calibration test verdict after bug-repair is WARN: the arm-shift parameter passes ($p = 0.166$) but the population-mean parameter is marginal ($p = 0.037$), and the repaired calibration covers only the two-arm submodel, not the production three-arm model. Convergence diagnostics here confirm chain mixing only and do not certify posterior calibration.

Table 24: Hierarchical Bayes Demand System Convergence Diagnostics (Extended Run)

Parameter	Mean	SD	95% CI	R-hat	ESS
ϕ_2	3.611	0.098	[3.424, 3.803]	1.005	5517
ϕ_3	4.062	0.098	[3.869, 4.252]	1.004	5658
bb_{foc}	-2.522	—	—	1.008	3866
bb_{price}	-1.195	—	—	1.000	10855

Note: 4 chains $\times$ 5000 iterations (warmup=2000), 3000 post-warmup draws per chain.

Extended from warmup=1000/draws=1000 per R-hat < 1.02 protocol. Joint Gibbs for $(\phi_2, \phi_3, \bar{\beta}_0)$.

Simulation-based calibration: $L = 50$ WARN ($p = 0.0103$); $L = 200$ FAIL (posterior under-dispersed; int

Direction-only corroboration: sign claims $P(\phi_2 > 0) = P(\phi_3 > 0) = 1.000$ are valid; interval widths not

C Funnel Scaling: Raw Nielsen HMS Aggregates

The funnel statistics in Section 4.7 are raw Nielsen HMS aggregate counts and projection-factor-weighted household summaries. Trial penetration (2.79%) and conditional repeat- ≥ 2 (39.1%) are computed directly from the Homescan purchase records without Berry–Levinsohn–Pakes (BLP) equilibrium calibration: trial penetration = trier households / footprint households; repeat rate = households with ≥ 2 purchase occasions / trier households.

A scaling factor Λ to map conjoint chip shares to trial appeal was explored in preliminary analyses using a BLP-style linear calibration to January 2012 yogurt scanner shares, following the ASSESSOR/BASES lineage [Silk and Urban, 1978]. This step was run as an exploratory robustness exercise and excluded as non-load-bearing: the funnel statistics reported in the paper are raw HMS aggregates computed directly from Homescan purchase records, with no BLP equilibrium calibration entering any reported figure. The descriptive validation does not depend on the BLP scaling step in any way.

D Worst-Case Dropout Bound Computation Details

Lee [2009] bounds apply the trimming argument to both tails of the return distribution under the monotone-compliance assumption: excess returnees in

the larger arm (Yoplait arm, $N = 368$) are drawn from one tail of the latent preference distribution. The trim fraction is computed as $1 - n_{\min}/n_{\max} = 1 - 282/368 = 23.4\%$ for the Yoplait versus Fruit arm comparison. The lower bound trims the top 23.4% of Yoplait arm share observations, and the upper bound trims the bottom 23.4%, yielding the one-sided worst-case bounds.

The primary share basis (defined in Section 2.2) is used for all bound computations; it ensures $\hat{l}b \leq \hat{\delta} \leq \hat{u}b$ by construction: the sample mean of trimmed observations lies within the range of all observations. The trim fraction of 23.4% is algebraically identical to the differential compliance implied by the equal-thirds invitation scenario (see Table 4), so the lower bound already absorbs the worst-case compliance scenario from that alternative.

Monotonicity is assumed throughout; it is untestable from the data. The most consequential violation risk is for category activation, where the comparable-return violation directly affects the outcome variable, reinforcing the directional-only classification.

E $\hat{\eta}^*$ Estimator: Large-Sample Properties

This appendix records the three formal claims for $\hat{\eta}^*(k) = \hat{\delta}_M(k)/\hat{\delta}_{\text{OSY}}(k)$ established in the math-audited estimator note.

Claim 1 (Consistency). $\hat{\eta}^*(k)$ is a consistent estimator of $\eta^*(k)$. Proof sketch: each arm mean $\bar{s}_b(k)$ converges almost surely to $E[s_b \mid \text{arm} = k]$ by the strong law of large numbers (bounded outcomes, i.i.d. within arm).

Joint convergence across arms follows from arm independence. The function $g(x, y) = x/y$ is continuous at $(\delta_M, \delta_{\text{OSY}})$ since $\delta_{\text{OSY}} > 0$ (empirically verified: $t > 12$). The continuous mapping theorem then gives $\hat{\eta}^*(k) \xrightarrow{a.s.} \eta^*(k)$.

Claim 2 (Asymptotic normality). Under i.i.d. sampling within each arm, as $\min(n_1, n_k) \rightarrow \infty$:

$$\sqrt{n}(\hat{\eta}^*(k) - \eta^*(k)) \rightarrow^d N(0, V(k)),$$

where $V(k)$ is the delta-method variance in Section 3.3. The central limit theorem applies because share outcomes are bounded; the delta method is valid because $\delta_{\text{OSY}} \neq 0$; the covariance terms account for the shared Yoplait arm observations entering both $\hat{\delta}_M(k)$ and $\hat{\delta}_{\text{OSY}}(k)$.

Claim 3 (Consistent plug-in variance). The plug-in variance estimator $\hat{V}(k)$, substituting sample variances and covariances for their population counterparts, is consistent for $V(k)$ by the continuous mapping theorem applied to consistent component estimators (sample second moments converge by the strong law of large numbers on bounded outcomes).

Numerical validation: delta-method and bootstrap SEs agree to within 2.5% (Table 11).

Sub-cell power note. At $N \approx 85\text{--}99$ per sub-cell, sub-cell bootstrap SEs are approximately $1.3\text{--}1.4\times$ arm-level SEs. Sub-cell $\hat{\eta}^*$ estimates address stability of the arm-level value, not sub-cell precision; individual sub-cell credible intervals should be interpreted with explicit width caveats.

References

- Mesfin Haile Alemu and Svein Ottar Olsen. An analysis of the impacts of tasting experience and peer effects on consumers' willingness to pay for novel foods. *Agribusiness*, 36(4):653–674, 2020.
- Kapil Bawa and Robert Shoemaker. The effects of free sample promotions on incremental brand sales. *Marketing Science*, 23(3):345–363, 2004.
- John T. Cacioppo and Richard E. Petty. Effects of message repetition on argument processing, recall, and persuasion. *Basic and Applied Social Psychology*, 10(1):3–12, 1989.
- Gary Chamberlain and Guido W. Imbens. Nonparametric applications of bayesian inference. *Journal of Business & Economic Statistics*, 21(1):12–18, 2003.
- Andrew T. Ching, Tülin Erdem, and Michael P. Keane. Learning models: An assessment of progress, challenges, and new developments. *Marketing Science*, 32(6):913–938, 2013.
- Min Ding, Rajdeep Grewal, and John Liechty. Incentive-aligned conjoint analysis. *Journal of Marketing Research*, 42(1):67–82, 2005.
- Tülin Erdem and Michael P. Keane. Decision-making under uncertainty: Capturing dynamic brand choice processes in turbulent consumer goods markets. *Marketing Science*, 15(1):1–20, 1996.

- Eleanor McDonnell Feit, Mark A. Beltramo, and Fred M. Feinberg. Reality check: Combining choice experiments with market data to estimate the importance of product attributes. *Management Science*, 56(5):785–800, 2010.
- Jens Hainmueller, Daniel J. Hopkins, and Teppei Yamamoto. Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments. *Political Analysis*, 22(1):1–30, 2014.
- Dominik Harz, Sebastian Hohenberg, and Christian Homburg. Virtual reality in new product development: Insights from prelaunch sales forecasting for durables. *Journal of Marketing*, 86(2):157–179, 2021.
- Linda F. Jamieson and Frank M. Bass. Adjusting stated intention measures to predict trial purchase of new products: A comparison of models and methods. *Journal of Marketing Research*, 26(3):336–345, 1989.
- David S. Lee. Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *Review of Economic Studies*, 76(3):1071–1102, 2009.
- John A. List and Craig A. Gallet. What experimental protocol influence disparities between actual and hypothetical stated values? *Environmental and Resource Economics*, 20(3):241–254, 2001.
- James J. Murphy, P. Geoffrey Allen, Thomas H. Stevens, and Darryl Weath-

- erhead. A meta-analysis of hypothetical bias in stated preference valuation. *Environmental and Resource Economics*, 30(3):313–325, 2005.
- Donald B. Rubin. The bayesian bootstrap. *The Annals of Statistics*, 9(1): 130–134, 1981.
- Jonas Schmidt and Tammo H. A. Bijmolt. Accurately measuring willingness to pay for consumer goods: A meta-analysis of the hypothetical bias. *Journal of the Academy of Marketing Science*, 48(3):499–518, 2020.
- Alvin J. Silk and Glen L. Urban. Pre-test-market evaluation of new packaged goods: A model and measurement methodology. *Journal of Marketing Research*, 15(2):171–191, 1978.
- Sean Talts, Michael Betancourt, Daniel Simpson, Aki Vehtari, and Andrew Gelman. Validating bayesian inference algorithms with simulation-based calibration. *arXiv preprint arXiv:1804.06788*, 2018.
- Glen L. Urban and Gerald M. Katz. Pre-test-market models: Validation and managerial implications. *Journal of Marketing Research*, 20(3):221–234, 1983.
- Robert B. Zajonc. Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology*, 9(2):1–27, 1968.