American Impact Review

Review Article

A Systematic Review of Variance Reduction Techniques in Online Controlled Experiments

Erik Silva*

Booking.com, Amsterdam, The Netherlands

* Corresponding author: Erik Silva, erik.silva@booking.com

OPEN ACCESS

Citation:
Erik Silva (2026). A Systematic Review of Variance Reduction Techniques in Online Controlled Experiments. *Am. Impact Rev.* 10.66308/air.e2026081

Received: July 24, 2026
Accepted: September 27, 2026
Published: September 28, 2026

DOI:
10.66308/air.e2026081

ISSN: 3071-124X

Abstract

Variance reduction in online controlled experiments encompasses changes to assignment, estimation, and the outcome being measured. This systematic review compares these mechanisms while retaining their estimands, information requirements, and inferential conditions. Exhaustively enumerated OpenAlex and arXiv searches to 25 September 2026, one backward citation pass, and bounded additional identification yielded 120 direct methodological works and 46 contextual sources, including three statistical foundations. Full-text extraction and retrospective descriptive coding mapped every main work to a primary method family, additional mechanisms, and forms of evidence. Covariate and prediction adjustment was the largest primary family, with 25 works; ratio, robust, distributional, and shrinkage estimation accounted for 19, and network or cluster design for 17. Analytical development appeared in 116 works, simulation or constructed experiments in 93, observed randomised contrasts in 56, and other empirical inputs in 51; these evidence categories overlap. The synthesis distinguishes exact fixed-coefficient identities from fitted-estimator asymptotics, target-preserving adjustment from metric construction, and variance from mean-squared error and detection frequency. Gains depend on assignment, treatment-effect heterogeneity, information timing, and uncertainty estimation. Incompatible baselines and targets preclude a pooled percentage benefit. Search restrictions, access barriers, and proprietary data limit the coverage and interpretation of the evidence.

Keywords: A/B testing, Covariate adjustment, Online controlled experiments, Statistical efficiency, Systematic review, Variance reduction

1. Introduction

Random assignment protects an experimental comparison from systematic allocation bias, but does not guarantee useful precision. Online outcomes differ markedly between users, repeated events are dependent, and the effect of an incremental product change may be small relative to this variation. More traffic or longer observation can improve precision, at the cost of scarce experimental capacity and slower decisions. Statistical variance reduction instead uses the assignment structure or additional information to estimate a specified effect more efficiently. This concern was already present in early guides to web experimentation (Kohavi et al., 2009).

The methods now described under this heading intervene at different stages. Stratification changes the comparisons created by assignment. Covariate adjustment removes predictable outcome variation after assignment. Triggering and policy-overlap methods use knowledge of where treatment could have changed outcomes. Other procedures stabilise a heavy-tailed response, borrow information across experiments, or construct a more sensitive proxy. These mechanisms cannot be assessed by the size of a reported percentage

alone. Metric decomposition, for example, can minimise variance or mean-squared error with different coefficients (Deng et al., 2024). A long-term proxy can detect the presence of a movement without estimating its magnitude (Zito et al., 2025).

Pre-experiment adjustment illustrates the distinction between an attractive principle and a transferable result. In the original CUPED study, queries per user benefited much more than revenue per user, while an inappropriate treatment-affected covariate reversed an estimated effect's direction (Deng et al., 2013). The covariance identity explains why prognostic information helps, but the experiment must also justify the covariate, the fitted coefficient, and the standard error. Classical work on regression adjustment shows why these are separate obligations: a conventional fitted adjustment need not improve precision, whereas an interacted estimator has a conditional asymptotic advantage under a specified randomisation scheme (Freedman, 2008b; Lin, 2013).

This review builds on existing syntheses rather than claiming an unreviewed topic. Larsen and colleagues reviewed transformations, control variates, and stratified sampling from a statistical perspective (Larsen et al., 2024). A broader systematic review examined the design and execution of A/B tests (Quin et al., 2023). Marketplace and continuous-experimentation reviews consider interference and organisational use of experiments, respectively (Auer et al., 2021; Bajari et al., 2023). Their differing scope motivates a comparison centred on the statistical target, required information, assignment, and evidential support of each precision claim.

The synthesis that follows is organised by three sources of precision: better experimental comparisons, better prediction of outcome variation that is irrelevant to the contrast, and stronger restrictions on the response being estimated. Design, network, and temporal methods act mainly through the first source; covariate, control-variate, and prediction adjustment through the second; heavy-tail treatment, trigger information, proxy construction, and model-based estimation increasingly through the third. Section 4 maps the corpus by method family within this frame, Figure 2 summarises it, and Section 6 returns to its implications.

The review synthesises 120 direct methodological works and 46 contextual sources. Every main work underwent full-text extraction, including works that supply an assumption, boundary case, or complementary mechanism rather than a headline numerical gain. The review first establishes its retrieval and coding procedure, then develops the statistical distinctions needed to interpret adjustment, maps the corpus across design and estimation, and compares evidence on its original scale. Its conclusions concern the accessible literature within these documented boundaries.

2. Review methods

2.1 Scope, information sources, and chronology

Eligible main works specified a statistical method that reduces estimator variance or improves precision in online randomised experiments, with an explicit connection to digital services, platforms, or products. Design, stratification, regression and prediction adjustment, complex outcomes, and other qualifying mechanisms were eligible. Faster computation of a standard error, sample-size prediction, or increased algorithmic reward alone did not establish this contribution. Methods allowing bias or changing the estimand were eligible when that change and its implications were identifiable. General statistical foundations, reviews, and adjacent inferential methods were retained separately as context.

A working protocol preceded pilot searching on 25 September 2026, but the review was not preregistered and an immutable initial version was not retained. Query versions and dated amendments record subsequent calibration. There was no lower date restriction or quota of included works. The upper boundary was 25 September 2026. Full-text methodological assessment was restricted to English because that was the language

supported by the reading and verification procedure.

OpenAlex Works and arXiv supplied the two systematic routes. Queries combined online-experimentation and A/B-testing terms with precision, variance reduction, covariate adjustment, design, prediction, complex metrics, and preceding reviews. Pilot unrestricted OpenAlex queries produced many indirect full-text or bibliographic matches; before substantive screening, its five final queries were restricted to the combined title-and-abstract field. Three final arXiv queries complemented them. The eight final query strings are reported verbatim, with their endpoints, search fields, and per-query retrieval totals, in Appendix B: the five OpenAlex queries returned 100, 411, 268, 68, and 128 occurrences, and the three arXiv queries returned 37, 36, and 12. Each result set was enumerated completely, with raw responses, exact parameters, timestamps, totals, and pagination states saved before filtering. Retrieved counts equalled the reported totals, and within-query uniqueness was checked. Crossref supported identity and metadata verification; no systematic search of Scopus, Web of Science, or the ACM interface is claimed.

2.2 Screening, versions, and the selection flow

OpenAlex returned 975 query occurrences representing 896 source records; arXiv returned 85 occurrences representing 66 source records. The resulting 1,060 occurrences and 962 records are retrieval units. Identifier and title matching, followed where necessary by author and content checks, grouped reports into works. The grouping rule was an exact DOI, arXiv-identifier, or verified-title correspondence; it merged 176 of the 962 records (18.3%) into multi-report work groups at this stage, leaving 786 screening candidates, and the grouped versions were retained rather than discarded. In the final map, after the four later groupings described below, 135 of 782 work groups contain more than one retrieved record; the full record-to-group correspondence is retained in the review records. Stable IDs and aliases preserved decisions when groups changed. Published reports, preprints, extensions, and accompanying data sets were distinguished rather than treated as interchangeable duplicate strings.

The historical abstract-screening snapshot contained 786 candidates: 640 excluded, 27 without adequate screening text, and 119 referred for full-text assessment. Four additional report-to-work groupings established during reading left 782 database works and 115 full-text candidates. The completed database branch contained 57 main works, 31 contextual works, 652 exclusions, and 42 unavailable records. The 42 comprise the 27 screening-text failures and 15 unavailable full texts. The other 12 full-text candidates were excluded substantively. Later groupings are shown at their actual position in Figure 1.

Term matching routed records for reading; it did not constitute final eligibility assessment. Decisions used an actual abstract or full text with a recorded reason. Missing and title-like abstracts prompted bounded recovery attempts. Full-text assessment required a complete document with verified identity; unavailability was kept separate from ineligibility. Statements extracted for synthesis required the full source and a physical-page locator. Selection used the version available by the cutoff, even where a later publication record existed.

One backward citation pass covered all 57 main database seeds, yielding 330 citation occurrences and 200 candidate titles. Twenty-seven matched existing records, and three were additional reports of other new candidates. The initial assessment retained 71 new works: 62 main and nine contextual. The remaining proposals comprised 11 unavailable texts or unresolved identities, seven scope exclusions, 13 adjacent topics recorded without extraction, and 68 general foundations not initially selected. Candidates were assessed from their own documents; the pass was not recursive. A separate bounded web route retained one main work and three predecessor reviews, with one unavailable candidate. These routes initially yielded 163 included works.

After the initial search and citation pass, a documented amendment added three statistical foundations: two previously identified contextual candidates and one distinct companion paper by Freedman. These three

general statistical sources do not enter the 120-work methodological denominator. The final inventory therefore contains 166 included works: 120 main and 46 contextual. All main works and 28 contextual sources have scientific extraction; 18 other context records document boundaries without extraction. Figure 1 distinguishes this addition from the original search and citation pass.

2.3 Extraction, descriptive coding, and appraisal

Extraction covered five dimensions: mechanism and comparator; metric, estimand, and units; auxiliary information and training; assumptions and uncertainty; and evidence provenance. Eight appraisal domains addressed target/baseline clarity, validity conditions, training/evaluation separation, simulation specification, empirical provenance and independence, Type-I error or coverage, code/data access, and transportability. Domain judgements retained reasons and applicability; they were not added into a quality score. Absence of a simulation, for example, is not an automatic defect in a theoretical work. The distribution of judgements across the eight domains is reported in Table 2, and per-work outcomes for all 120 main works are listed in Appendix A.

Screening, extraction, descriptive coding, and the verification checks were performed by computational agents; there was no independent duplicate human screening or extraction, and inter-reviewer agreement was not measured. All 120 main records were coded from full-text extraction and page-level evidence, with ambiguous cases checked against the original PDFs. The classification was retrospective: coding definitions were specified after the initial review and before aggregation. Each work received one primary family for its central mechanism and optional secondary families for substantive additional mechanisms. The nine primary categories are mutually exclusive; secondary labels overlap. Mere use of switchback data does not establish a contribution to switchback design, and a temporal setting alone does not imply model-based dynamic estimation.

Four overlapping indicators recorded analytical argument or derivation, simulation or constructed experiments, observed randomised contrasts, and other empirical inputs. Synthetic effects imposed on real covariates count as constructed evidence; the real inputs are separately recorded. Observed A/A outcomes qualify as randomised contrasts, but simulated assignments do not. Analytical coding does not imply a new theorem. Counts use works as their denominator, not experiments or independent platforms. Each code was linked to its source evidence and interpreted as an analytical contribution, empirical finding, or boundary on applicability.

Publication status was audited through retained and bounded exact-item publisher, DOI, and repository evidence. Work-level publication, the inspected file and date, the earliest evidenced full report, and evidence of peer review were recorded separately. A DOI alone establishes neither peer review nor equivalence of manuscript versions. The year profile retains unknown dates and gives its convention explicitly. This verification did not constitute another search for eligible studies. The bibliography uses confirmed canonical publications; extraction and page locators remain tied to the inspected versions, including expanded reports.

The synthesis did not include new experiments, comprehensive proof replication, or a quantitative meta-analysis. The complete included-work list with per-work appraisal outcomes is reported in Appendix A, and the query specifications in Appendix B, rather than being available only on request.

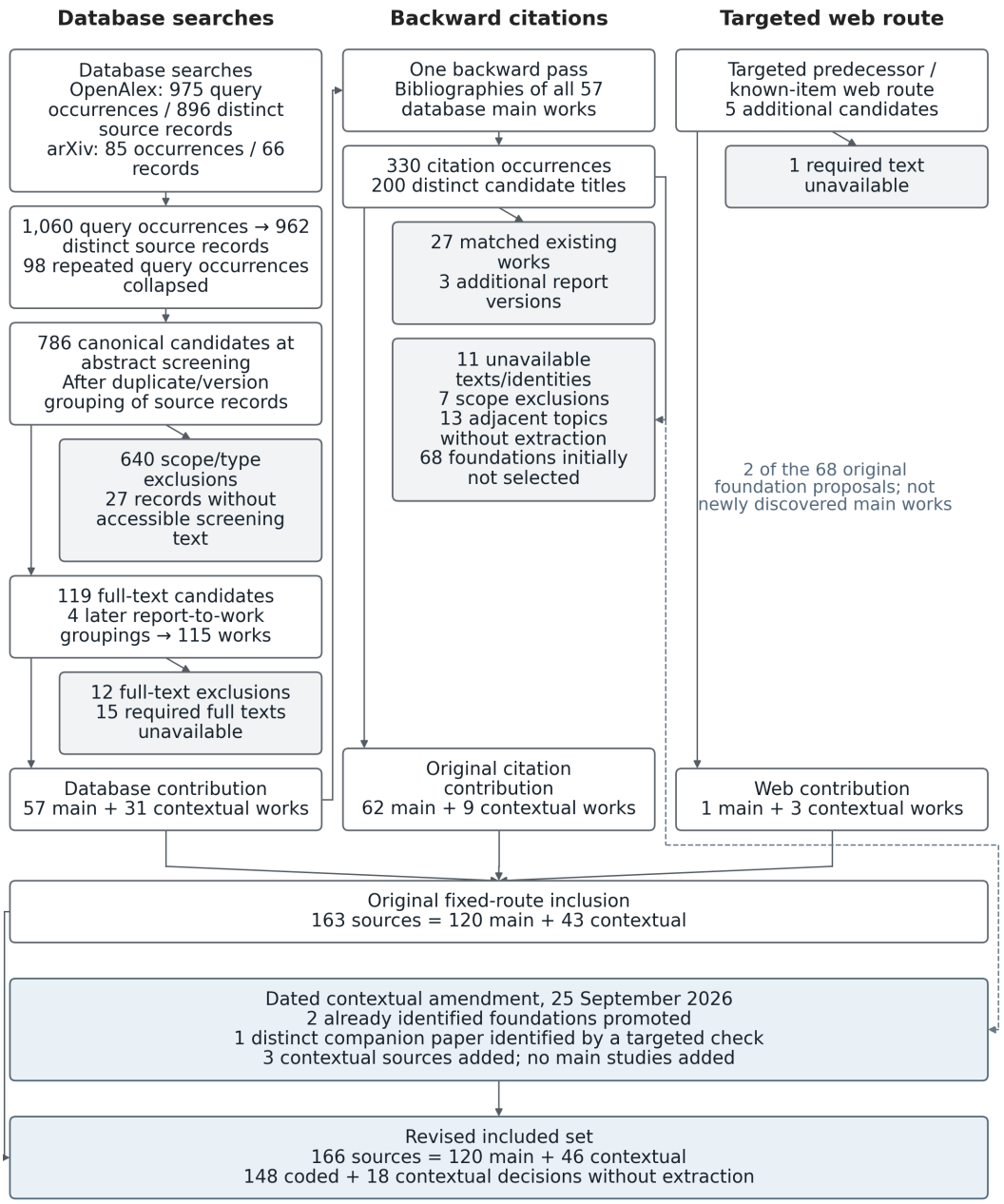

Figure 1. Recorded selection and version-grouping flow. Query occurrences, source records, historical screening candidates, and final work groups are different units. The foundation amendment adds three contextual sources after the original 163-work inventory; two were already identified in the citation pass and one is a distinct companion paper.

3. Statistical foundations

3.1 Targets and measures of improvement

Precision is defined relative to an estimand. Let a finite experimental population have potential outcomes $Y_i(1)$ and $Y_i(0)$, and let $\tau = n^{-1} \sum_i (Y_i(1) - Y_i(0))$ be its mean treatment effect. A superpopulation target averages over another source of randomness; a conditional target holds additional information fixed. The point-

estimator algebra alone does not choose among these interpretations. Arm-specific regression constructions can coincide numerically while requiring different uncertainty calculations for different targets (Zhang et al., 2025b).

For any estimator, $\mathrm{MSE}(\hat{\tau}) = \mathrm{Var}(\hat{\tau}) + \{E(\hat{\tau}) - \tau\}^2$. This decomposition separates variance reduction from an improvement that trades variance against bias. Metric decomposition supplies different coefficients for these objectives (Deng et al., 2024). Shrinkage can improve average accuracy over many effects without delivering unbiased estimation or calibrated intervals for every individual treatment (Dimmery et al., 2019). Neither objective is intrinsically illegitimate; the comparison must state which one is pursued.

Changing the outcome also changes what a precision gain means. A trimmed mean, a weighted mean under heterogeneous effects, and a proxy for a future effect need not estimate the original average effect. Standard-error reductions and root mean-squared error reductions additionally use different scales: a standard-error ratio s corresponds to a variance ratio s^2, whereas the analogous squared root-MSE ratio includes bias. Detection frequency also depends on effect sizes, the decision rule, and the metric. We preserve these distinctions rather than treating every reported improvement as a variance ratio.

3.2 Fixed controls and fitted regression adjustment

For fixed θ and fixed reference μ_X, define $Y^\star = Y - \theta(X - \mu_X)$. Then

$$\mathrm{Var}(Y^\star) = \mathrm{Var}(Y) + \theta^2 \mathrm{Var}(X) - 2\theta \mathrm{Cov}(Y, X).$$

If $\mathrm{Var}(X) > 0$, the optimal fixed coefficient is $\theta^\star = \mathrm{Cov}(Y, X)/\mathrm{Var}(X)$, giving $\mathrm{Var}(Y^\star) = \mathrm{Var}(Y)(1-\rho^2)$. CUPED uses this control-variate principle with prognostic pre-experiment information (Deng et al., 2013). The identity assumes a fixed reference; an estimated common centre either cancels in the contrast or requires its randomness to be handled. It also says nothing by itself about a coefficient selected from the same outcomes being analysed.

To make that distinction explicit, suppose exactly n_1 of n units receive treatment under complete randomisation, with $n_0 = n - n_1$. Potential outcomes and unaffected covariates X_i are fixed. Write $x_i = X_i - \overline{X}$, where $\overline{X}$ uses the full experimental population. Separate intercept-containing regressions within the arms give slopes $\hat{b}_1$ and $\hat{b}_0$ and the estimator

$$\hat{\tau}_{\text{int}} = \{\overline{Y}_1 + (\overline{X} - \overline{X}_1)^\top \hat{b}_1\} - \{\overline{Y}_0 + (\overline{X} - \overline{X}_0)^\top \hat{b}_0\}.$$

This is also the treatment coefficient in a regression on $1, Z_i, x_i, Z_i x_i$. Both fitted means are evaluated at the same covariate mean. A common pooled slope removes the interactions and gives conventional adjustment. Neither construction requires the linear regression to be the true outcome model (Lin, 2013).

If instead the slopes b_1, b_0 are fixed before assignment, put $r_i(z) = Y_i(z) - x_i^\top b_z$. With finite-population variances using divisor $n - 1$, randomisation gives

$$E_Z\{\hat{\tau}(b_1, b_0)\} = \tau, \quad \mathrm{Var}_Z\{\hat{\tau}(b_1, b_0)\} = \frac{S^2_{r(1)}}{n_1} + \frac{S^2_{r(0)}}{n_0} - \frac{S^2_{r(1)-r(0)}}{n}.$$

This is an algebraic application of the finite-population mean and covariance identities to residualised outcomes (Freedman, 2008a). It is exact for fixed slopes. Substituting in-experiment fitted slopes does not preserve its exact unbiasedness claim. Fitting creates dependence between accidental covariate imbalance and the

correction; under the relevant regularity conditions its leading bias can be of order n^{-1}, smaller than the standard-error scale $n^{-1/2}$, but not identically zero (Lin, 2013).

3.3 What the asymptotic guarantee does and does not establish

Freedman demonstrated that conventional common-slope adjustment can increase asymptotic variance and that its usual homoskedastic standard error can mislead (Freedman, 2008b). Lin's interacted estimator resolves a specific part of this problem. His result assumes complete two-arm randomisation, a fixed covariate dimension, bounded fourth moments, convergent required moments, a nonsingular limiting covariate Gram matrix, and $n_1/n \to p \in (0,1)$ (Lin, 2013). These are conditions on the actual assignment units; many events from a few users or clusters do not create an equivalent asymptotic sequence.

Let q_z be the limiting population linear-projection slope for outcome $Y_i(z)$, and let V denote the limiting variance of $\sqrt{n}(\hat{\tau} - \tau)$. Writing Var_N for finite-population variance with divisor n, the asymptotic gain is

$$V_{\text{unadj}} - V_{\text{int}} = \frac{\lim_{n\to\infty} \mathrm{Var}_N[x_i^\top \{(1-p)q_1 + pq_0\}]}{p(1-p)} \geq 0.$$

The result accommodates nonlinear outcome relations and heterogeneous effects. It concerns a particular weighted combination of outcome slopes: strong prediction of both arms need not give a strict gain if the relevant slopes cancel. The expression is on the $\sqrt{n}$ scale; the corresponding first-order difference for the unscaled estimators is divided by n (Lin, 2013).

The cost of omitting interactions is separately characterised by

$$V_{\text{pool}} - V_{\text{int}} = \frac{(2p-1)^2}{p(1-p)} \lim_{n\to\infty} \mathrm{Var}_N[x_i^\top (q_1 - q_0)] \geq 0.$$

It disappears at equal allocation or equal population projection slopes. Thus unequal allocation and effect-related slope differences provide a concrete reason for interactions; common-slope adjustment is not uniformly inferior to the unadjusted estimator in every setting. Nor does balanced allocation alone settle the multi-arm problem: a third arm can influence a pooled slope used for a pairwise contrast, as Freedman's counterexample demonstrates (Freedman, 2008a).

Inference remains a separate calculation. Under Lin's conditions, the Huber–White estimate for the interacted contrast is consistent or conservative; unexplained individual-effect heterogeneity contributes to asymptotic conservatism. Small samples and high leverage can still compromise the approximation (Lin, 2013). These results provide neither a generic cluster-robust variance formula nor a guarantee for interference, switchbacks, adaptive allocation, increasing-dimensional prediction, or ratios of random totals. Those settings require their own assignment, fitting, and uncertainty arguments. This boundary connects the general foundations to the specialised methods below.

4. Corpus profile and methods across design and estimation

This section profiles the corpus and then follows the three sources of precision stated in Section 1. Allocation, network, and temporal designs (Section 4.2) act on the comparisons the experiment creates; covariate, prediction, and in-experiment adjustment (Sections 4.3 and 4.5) predict outcome variation irrelevant to the contrast; heavy-tailed and distributional estimation (Section 4.4), the trigger information and proxy construction within Section 4.5, and policy structure and dynamic modelling (Section 4.6) increasingly restrict

or change the response being estimated, and sequential procedures (Section 4.7) alter the decision rule around the estimate. Most works combine sources, which the family coding records through secondary labels.

4.1 Composition of the main corpus

Table 1 profiles all 120 main work groups. Covariate and prediction adjustment is the largest primary family (25), followed by difficult-outcome estimation (19) and network or cluster design (17). The nine exclusive families organise the corpus; the discussion combines related families where the statistical comparison is more informative than a one-family-per-subsection structure. With secondary labels included, covariate adjustment appears in 50 works, illustrating how often it accompanies a design or specialised metric rather than constituting the entire method.

Analytical development is reported in 116 works, constructed or simulation evidence in 93, observed randomised contrasts in 56, and other empirical inputs in 51. These overlapping counts show the mixture of support, not its independence or strength. Three works are coded as analytical-only, whereas 40 combine analytical development, simulation, and other empirical inputs without observed randomised contrasts. A simulator calibrated to platform logs offers a different test from an observed treatment comparison, even when both use large data sets.

By 25 September 2026, 86 work groups had a publication record linked to the work, and one had a confirmed publication of an earlier report only. Twenty-nine had no confirmed publication, and four remained unresolved. These counts include journal, conference, and workshop records; they do not establish that every inspected version or expanded result was published or peer reviewed. In particular, the expanded switchback report retains its separate version status despite its published predecessor (Xiong et al., 2024).

The year groups in Table 1 use the earliest evidenced full report among retained records, with one undated work retained in the denominator. Nineteen works first appear in that evidence in 2026, whereas 24 inspected files are dated 2026 and ten work groups have a confirmed publication dated 2026. These measures distinguish the arrival of a report, the revision used for extraction, and publication. They are not interchangeable counts of new studies, and 2026 covers only the period through the cutoff. The triple matters for interpretation: for 25 of the 120 works the inspected version postdates the earliest evidenced full report, and among the 87 groups with a linked publication record, 27 have a publication year that differs from the first-report year. A result attributed to a year under one convention can therefore move under another, and Section 6 uses the same distinction when interpreting families that depend on unpublished or expanded reports.

Table 1. Descriptive profile of 120 main work groups. Primary families, year groups, and publication strata each sum to 120; evidence indicators overlap. Percentages use 120 as the denominator and describe works, not independent experiments.

Dimension	Category	Works, n	%
Primary family	General allocation, balance, and stratification	10	8.3
Primary family	Network and cluster design	17	14.2
Primary family	Switchback and panel design	11	9.2
Primary family	Covariate, control-variate, and prediction adjustment	25	20.8
Primary family	Ratio, robust, distributional, and shrinkage estimation	19	15.8
Primary family	Metric construction and proxies	9	7.5
Primary family	Trigger, exposure, and policy information	15	12.5
Primary family	Dynamic and dependence modelling	11	9.2
Primary family	Sequential, multistage, and testing procedures	3	2.5
Evidence indicator	Analytical argument or derivation	116	96.7
Evidence indicator	Simulation or constructed experiment	93	77.5
Evidence indicator	Observed randomised-experiment contrasts	56	46.7

Continued on next page

Dimension	Category	Works, n	%
Evidence indicator	Other empirical data	51	42.5
Earliest evidenced full-report year	2019 or earlier	34	28.3
Earliest evidenced full-report year	2020–2022	24	20.0
Earliest evidenced full-report year	2023–2024	31	25.8
Earliest evidenced full-report year	2025	11	9.2
Earliest evidenced full-report year	2026, through 25 September	19	15.8
Earliest evidenced full-report year	Unknown	1	0.8
Publication status at cutoff	Publication record linked to work	86	71.7
Publication status at cutoff	Published earlier report only	1	0.8
Publication status at cutoff	Repository/report; publication not confirmed	29	24.2
Publication status at cutoff	Publication unresolved	4	3.3

Year groups describe the earliest full-report date established in retained evidence, not necessarily the first circulation of a work. Publication strata concern linked work-group records and retain the distinction between an earlier report and the inspected version. Unknown dates and unresolved status remain in the denominator.

Table 2 reports the distribution of the eight appraisal domains across the 120 main works; per-work judgements appear in Appendix A. The domains record reporting and validity conditions, not a quality score, and "not applicable" preserves works for which a domain does not arise, such as simulation description in a purely analytical paper. Three patterns informed the synthesis. Transportability is a concern for all 120 works, reflecting single-platform, proprietary, or domain-specific settings; this motivates the refusal in Section 5 to pool gains across sources. Validity assumptions raise concerns in 93 works and empirical provenance or independence in 66, which supports the emphasis on conditional guarantees in Section 5.2. Code or data availability is unclear in 89 works, so most reported gains cannot currently be recomputed from released materials.

Table 2. Appraisal outcomes across eight domains for the 120 main works. Rows sum to 120. Judgements describe what each work reports, retain applicability, and were not combined into a score.

Appraisal domain	Reported adequate	Concern	Unclear	Not applicable
Target and baseline clarity	94	25	1	0
Validity assumptions	23	93	4	0
Training and evaluation separation	40	27	25	28
Simulation description	62	29	3	26
Empirical provenance and independence	9	66	17	28
Type-I error and coverage	24	27	68	1
Code and data availability	25	4	89	2
Transportability	0	120	0	0

4.2 Allocation, stratification, and experimental structure

Design-based methods reduce variation by controlling which units are compared before outcomes are observed. Stratification creates comparisons within groups expected to have similar outcomes, while matching or rerandomisation rejects allocations with poor balance. Time creates a related problem in continuously arriving populations: users assigned at different moments can be exposed to different background conditions. Work on non-stationary A/B tests combines time poststratification with locally balanced assignment across arrivals. Its efficiency claims depend on restrictions on the growth of the number of strata and on the underlying arrival, smoothness, and noise assumptions (Wu et al., 2022).

Sequential allocation can also use observed baseline covariates without waiting for the entire experimental population. The pigeonhole design places arriving users into predefined covariate cells and balances assignments within those cells. Its procedural advantage is the ability to act on information available at arrival; it does not establish a universal gain under every possible outcome model. The method requires access to the assignment process, suitable baseline measurements, and the design information assumed by its analysis. These requirements distinguish it from an analysis-stage correction that can be applied after an existing experiment has finished (Zhao and Zhou, 2025).

Network experiments make the objective of balance more complicated. Randomising connected units together can reduce contamination, yet larger correlated groups can increase sampling variability. A marketplace simulation of matched graph-cluster assignment illustrates that reduced bias can coexist with a wider sampling distribution than individual assignment. Rerandomisation criteria for network experiments likewise depend on the chosen outcome model and causal target. The relevant design decision is therefore a bias–variance comparison under a stated interference model, not a rule that more clustering necessarily improves precision (Holtz and Aral, 2020; Zhang, 2025).

Temporal designs require an equally explicit treatment of carryover. Rerandomisation in a panel can reduce the variance associated with prognostic imbalance under a no-carryover approximation. The same analysis shows that longer-lasting treatment effects can make bias dominate error even when variance falls. A design optimised for contemporaneous contrasts may therefore be inappropriate for a cumulative or persistent effect. Where carryover cannot be ruled out, its treatment belongs in the estimand and model specification before the balance criterion is chosen (Zeng et al., 2026).

Switchback work develops two further responses to persistence. A minimax schedule controls worst-case uncertainty for a specified persistent-exposure effect under bounded potential outcomes (Bojinov et al., 2023). Data-driven designs instead use historical event density and carryover curves to choose schedules through empirical-Bayes mean-squared error (Xiong et al., 2024). The first protects against a stated class of potential outcomes; the second spends modelling assumptions to adapt to the observed environment. Under geometric mixing, a corrected estimator restores observations when neighbouring blocks retain the same treatment, whereas plain block burn-in discards observations after every boundary (Hu and Wager, 2026). These distinctions concern assignment history and carryover, not simply another covariate in a fixed experiment.

Implementation can materially change the benefit of an otherwise familiar design. Netflix's comparison found similar performance for CUPED and poststratification when they used the same covariates, while a distributed stratified assignment achieved less variance reduction than its single-machine counterpart for existing users. This evidence does not discredit stratification; it shows that the information available to an implementation and the way it coordinates assignments matter. A useful comparison should hold those features constant where possible. Table 3 places allocation methods alongside estimation methods to make their different intervention points and resource requirements explicit (Xie and Aurisset, 2016).

Table 3. Mechanisms, information requirements, and conditions for interpreting precision gains.

Method family	Information or intervention	Target and principal condition	Illustrative sources
Allocation and stratification	Baseline groups, sequential balance, or network-aware assignment	A specified experimental contrast; balance and the assignment mechanism must enter the analysis	(Xie and Aurisset, 2016; Zhang, 2025; Zhao and Zhou, 2025)
Linear control variates	Historical outcomes or fixed baseline features	The original mean contrast if the adjustment has zero treatment contrast; estimated coefficients require justification	(Deng et al., 2013; Zhang et al., 2025b)
Prediction-assisted adjustment	External, out-of-fold, or leave-one-out predictions	The stated mean effect under the corresponding assignment and training conditions; prediction quality governs the opportunity	(Gagnon-Bartsch et al., 2023; Guo et al., 2021; Sales et al., 2023)
Ratio and distributional methods	Numerator–denominator structure, tail, or outcome-distribution models	A ratio, quantile, trimmed mean, or model-restricted effect; these targets are not interchangeable	(Athey et al., 2023; Baweja et al., 2024; Charette and Boudreault, 2025; Lopes et al., 2019)
Zero-effect augmentation	Trigger complements or selected in-experiment components	The overall effect only when the augmentation has mean zero under the required response assumptions	(Deng et al., 2023b; Deng and Hu, 2015; Lin and Crespo, 2026)
Policy and dynamic structure	Known policy overlap, state transitions, or treatment locality	A policy-value or time-horizon effect under overlap, model, and dependence assumptions	(Chen et al., 2025; Chihara et al., 2026; Jeunen, 2026)
Metric learning and shrinkage	Past experiments, multiple outcomes, or multiple treatment arms	A learned proxy or a potentially biased estimate; assess target alignment and mean-squared error separately	(Dimmery et al., 2019; Tripuraneni et al., 2024; Zito et al., 2025)

4.3 Linear and machine-learning adjustment

Pre-experiment adjustment is attractive because historical measurements can be prognostic without lying downstream of the experimental treatment. The strongest gains arise when those measurements predict the outcome relevant to the present test, rather than merely describing the user in general. The original CUPED evidence is informative precisely because it contains uneven results: revenue per user showed less than 5% reported variance reduction, despite much larger reductions for engagement measures. The covariance identity explains this variation without requiring a different statistical objective for each metric. It also suggests why absence of a strong historical predictor can limit additional benefit (Deng et al., 2013).

Machine learning can capture nonlinear and interactive predictive structure that a small linear model misses. In the Yandex study of boosted-tree adjustment, the mean variance ratio for the sessions metric was 0.3734 relative to the unadjusted estimator across 161 experiments. This is a substantial within-source result, but it is an average of reported ratios for one metric and platform, not a transportable review-level effect. The study's local predictors were trained on the same experimental users used for effect estimation. That implementation

fact should be distinguished from a general validity guarantee: it motivates scrutiny of how flexible fitting interacts with randomisation (Poyarkov et al., 2016).

Cross-fitting addresses this interaction by generating each unit's prediction from a model trained without its own analysis fold. MLRATE combines such out-of-fold predictions with treatment-interacted ordinary least squares. Under its stated assumptions, its asymptotic variance is no greater than that of the unadjusted difference in means, and the paper provides simulation evidence about interval coverage. The comparison is with that baseline, not with every competing adjustment, and an asymptotic statement does not promise that every realised small experiment will improve. Cross-fitting should be understood as part of the statistical construction, not simply a software optimisation (Guo et al., 2021).

Under independent Bernoulli assignments, assignment-independent potential-outcome imputations yield an exactly unbiased residualised inverse-probability estimator (Gagnon-Bartsch et al., 2023). ReLOOP incorporates predictions from auxiliary data into a leave-one-out analysis. Its educational evaluation supplied 227 contrasts from 68 A/B tests, with shared arms limiting the independence of those contrasts (Sales et al., 2023). Compared with the fixed-dimensional interacted regression in Section 3, these methods make the assignment independence of the fitted contribution an explicit part of their construction. The appropriate comparison is therefore between complete training-and-estimation procedures.

External data can be valuable when the experiment itself has limited predictive training information, but more elaborate prediction is not automatically beneficial. A poor predictor can worsen precision, as illustrated in the ordered LOOP evaluation. Calibrated approaches can estimate how strongly predictions should influence the outcome contrast, rather than imposing a unit coefficient. The arm-specific calibrated estimator examined in the corpus uses fitted slopes for the difference between population-average and observed-arm predictions. Its no-harm interpretation remains asymptotic and conditional on its specified construction, rather than an exact guarantee for arbitrary model fitting (Arbour et al., 2026; Pei et al., 2026).

Past experiments can also inform the statistical model itself. A family-learning approach estimates a local exponential-family structure from related historical groups, then uses a learned score in later comparisons. Its efficiency argument depends on small local effects and regularity of the learned family; a score contrast need not equal the original raw-mean effect outside that model. Independently learned scores can nevertheless be used with held-out permutation testing for null assessment. This separates two benefits that are sometimes conflated: learning a statistic suited to a particular alternative and preserving the validity of the test used to evaluate it (Fithian and Ting, 2017).

4.4 Ratio, heavy-tailed, and distributional metrics

Ratio metrics require attention to both components. A user-level linearisation subtracts the denominator multiplied by an appropriate control-arm ratio from the numerator, allowing adjustment at the randomisation-unit level. More general procedures model numerator and denominator outcomes separately within each treatment arm and then reconstruct the target ratio. Their efficiency depends on nuisance-regression quality and their validity depends on the denominator assumptions used. In one reported simulation, imposing denominator stability when it was false produced zero reported coverage for nominal 95% intervals in both simulated dimensions. That result warns against using a simpler ratio target merely because it is computationally convenient (Baweja et al., 2024; Jin and Ba, 2023).

Earlier ratio-transformation evidence makes the target distinction explicit. With positive denominators and the specified control-arm coefficient, a linearised contrast preserves the sign of a ratio effect, while its magnitude is rescaled. Approximate significance equivalence requires further conditions, including restrictions on denominator changes under the null. The Yandex evaluation included 390 tests, with a 197-test subset used

for regression adjustment. These sample counts describe comparisons within one source; they do not remove the conditions needed to interpret the transformed metric or make it numerically identical to the ratio effect (Budylin et al., 2018).

Clustered ratios combine two issues that should not be solved independently. The double-residual calculation in the clustered power-analysis study accounts for covariate adjustment and the covariance between numerator and denominator at the assignment-unit level. Its planning examples illustrate how specifications can change projected power, but projected power is not an observed rate of successful future experiments. Similarly, its large empirical collection concerns a particular customer-level outcome and should not be conflated with a separate per-order worked example. Keeping the metric and unit attached to each result prevents apparently similar percentages from being compared on incompatible foundations (Hesterberg and Knight, 2024).

Heavy-tailed methods reveal a substantive choice between changing the target and modelling its difficult part. Trimming can remove precisely the tail in which treatment acts, changing the detected effect (Charette and Boudreault, 2025). By contrast, a semiparametric Bayesian estimator retains a mean target by combining observed lower outcomes with a modelled generalised-Pareto tail mean; its reported root mean-squared error includes model bias (Lopes et al., 2019). When rare large outcomes matter economically, discarding and modelling them address different scientific problems, despite both producing a more stable estimate.

Stratification can be combined with such preprocessing. A monetisation-metric procedure weights within-stratum CUPED-adjusted effects by population stratum shares, but its implementation also trims or winsorises the observed outcome. Historical strata and outcome-tail removal serve different purposes; the latter prevents the resulting gain from being attributed automatically to improved estimation of the untrimmed mean effect (Pokharna et al., 2026).

A related approach borrows the tail distribution from a much larger background sample and reweights it through a bounded exponential tilt. It targets a population mean but assesses a potentially biased estimator through both variance and mean-squared error. Shared tail behaviour, background-sample growth, and the threshold choice are substantive assumptions. When the population variance is infinite, a finite-sample comparison of empirical errors is particularly unsuitable for casual translation into a universal percentage variance reduction. The source's Facebook-based resampling experiment supplies evidence about the observed distributions, rather than independent replications across deployments (Fithian and Wager, 2015).

Distributional targets require another distinction between improving an estimator and improving its uncertainty calculation. Conditional-distribution adjustment can subtract an arm-average predicted distribution function and restore the average prediction across all randomised units. The distributional-estimation paper's central theorem is expressed using oracle conditional probabilities, so learned-model performance needs separate support. By contrast, a method that estimates quantile uncertainty from a cluster-level distribution-function variance, or balances bootstrap draws to reduce Monte Carlo noise, may leave the underlying quantile-effect estimator unchanged. Those contributions belong in the surrounding inferential context but should not be counted as demonstrated reductions in its sampling variance (Hirata et al., 2025; Liu et al., 2018; Wang and Zhang, 2021).

4.5 Treatment-insensitive components and in-experiment information

Historical covariates are not the only possible source of a valid control contrast. One-sided triggering creates groups known not to receive the active treatment, and their outcomes can contribute a mean-zero augmentation to the overall intention-to-treat estimate. The corresponding method uses weighting to connect unexposed units to the relevant control population. Its strongest reported company-example variance reduction failed the method's own mean-zero diagnostic, while simulation showed bias when an important triggering–outcome

confounder was omitted. The evidence therefore supports a qualified augmentation strategy, with its diagnostic and identifying assumptions carried alongside its precision gain (Deng et al., 2023b).

Advertising designs can construct counterfactual exposure information through control advertisements or logged ghost opportunities. This can support removal of unexposed users or outcomes preceding the first eligible exposure, provided the counterfactual opportunity is defined symmetrically. The resulting exposed-user or local effect must be distinguished from the all-user intention-to-treat effect. Conditioning instead on later retargeted exposure frequency can invalidate that symmetry. Precision therefore depends on how exposure is generated and recorded, as well as on the final statistical adjustment (Johnson et al., 2017b, 2017a).

Trigger-complement adjustment and combinations of pre-period and in-period data develop related ideas. An all-user contrast can be corrected with variables whose causal effects are assumed to be zero, but this requires more than observing a small estimated difference between arms. The Etsy study treats non-rejection of balance as a diagnostic and records a broad feature set across 29 experiments; that does not establish universal treatment-insensitivity or independent validation of feature selection. Estimating correction coefficients and selecting apparently balanced features from the same data introduce further uncertainty. The appropriate safeguard is an explicit causal and statistical construction, followed by checks sensitive to its likely failure modes (Deng and Hu, 2015; Lin and Crespo, 2026).

Outcome prediction during an experiment can instead target sparse or delayed measurements. Under a suitable surrogacy condition, replacing an outcome by its true conditional expectation can preserve the relevant treatment contrast. That identity does not automatically survive an estimated predictor, capping, a treatment pathway omitted from the conditioning variables, or selection based on an in-period exposure score. The study of efficient targeted measurement shows that quantile-based selection must be included in uncertainty estimation. Forecasting an outcome and adjusting an effect estimate are therefore different operations unless their equivalence is established for the target and data-generating assumptions (Deng et al., 2023a).

Proxy-metric development makes this distinction especially visible. An optimal composite proxy learned from previous experiments can depend on the noise level, and therefore the sample size, of the experiment in which it will be used. A more sensitive proxy may anticipate the existence of a long-term effect without estimating its magnitude. Future-engagement prediction similarly changes the analysed metric when its arm means are compared directly. Such work can improve a decision system, but it should be described as metric or target construction, with evidence about transportability across interventions, rather than as an automatic improvement of the original unbiased effect estimate (Drutsa et al., 2015; Tripuraneni et al., 2024; Zito et al., 2025).

Evaluation across held-out experiments can test that transportability. One learned-metric study fits its weights leaving out the evaluated experiment, then assesses its resulting test statistic. This separates weight fitting from evaluation; the source does not establish that selection of the input metrics was also repeated inside every training fold. The distinction matters when assessing the whole metric-development procedure (Jeunen and Ustimenko, 2024).

4.6 Policy structure, dynamic outcomes, and borrowing information

Policy comparisons can exploit structure unavailable to an ordinary two-sample mean. If two policies assign the same probability to an action, a counterfactual contrast can give that action zero difference weight and remove its reward noise. The argument requires known logging probabilities, sufficient support, and a reward model that is invariant to policy conditional on the relevant information. Logging-policy optimisation makes the variance of a counterfactual click-through-rate difference an explicit objective. Neither overlap alone nor a zero-weight identity establishes variance dominance under arbitrary logging, feedback, or interference

(Jeunen, 2026; Oosterhuis and Rijke, 2020).

Item-level decompositions provide another way to share information: an expected post-click outcome can be expressed through item click probabilities and conditional item outcomes. This requires post-click behaviour to be independent of ranking conditional on the item, together with an adequate click model. In the reported real-service application, an additional feature-based variance predictor did not improve efficiency, despite the broader method's simulation motivation (Iizuka et al., 2021).

Ranking applications illustrate why the evaluation metric must remain visible. Airbnb's counterfactual comparison downweights bookings when their positions are similar under the two rankings, changing the evaluation metric. Learned scoring can also reverse a preference: the extended Yahoo evaluation reported reversals in four of 71 comparisons. Interleaving and multileaving methods are valuable neighbouring approaches, but some optimise entropy surrogates or preference-error criteria rather than effect-estimator variance. A simulated preference-error improvement, even a large one, cannot therefore be translated into an equal percentage reduction in A/B sampling variance (Brost et al., 2016; Chapelle et al., 2012; Radlinski and Craswell, 2013; Zhang et al., 2025a).

Dynamic methods account for the way treatment changes future states as well as immediate outcomes. A treatment-locality model can share information from states whose transitions or rewards are known to be unaffected, while a dynamic adjustment can combine residual corrections with treatment-specific history transitions. These gains rely on the locality or transition assumptions rather than on prediction alone. The information-sharing bound in the multi-arm study applies to a homogeneous-state component of asymptotic variance, and misspecifying the affected states can introduce bias. Similarly, a sequential optimisation surrogate that drops within-trajectory covariance is not automatically an optimiser of the full variance expression (Chen et al., 2025; Chihara et al., 2026; Sakhi et al., 2026).

Borrowing information across effects also motivates shrinkage. A normal-prior Bayesian comparison shrinks the observed effect and its conditional posterior variance using the same prior-dependent factor. This is a conditional posterior statement, not a guarantee of lower frequentist error for every treatment. The online-shrinkage study found that aggregate accuracy gains did not preserve nominal interval coverage for every arm. These methods are relevant when the decision objective is repeated-experiment accuracy, but their use should make the prior, the population of exchangeable effects, and the cost of individual-effect bias explicit (Deng, 2015; Dimmery et al., 2019).

4.7 Sequential and multistage procedures

The smallest primary family contains three works whose efficiency problem includes the decision sequence. Design-based panel confidence sequences use a proxy fixed before the current assignment to reduce a conservative variance bound; their Netflix reanalyses illustrate sequential inference rather than measuring an adjusted-variance gain (Ham et al., 2026). Two-stage selection and validation instead target a selected treatment, combining information while correcting selection bias through regression or shrinkage (Deng et al., 2014). Empirical-Bayes testing learns the distribution of effects across experiments and produces prior-dependent effect summaries (Deng, 2015). These procedures operate on different targets and error criteria. Their common contribution is to include the selection or stopping rule in the statistical object being improved, rather than append it to an otherwise fixed-sample comparison.

Figure 2 summarises the organising interpretation used here: methods can act on allocation, remove predictable outcome variation, exploit a restricted treatment response, or change the target itself. These categories overlap; for example, a ratio estimator may combine a design intervention, historical prediction, and a treatment-insensitive component. The diagram is a conceptual synthesis, not a fitted model or an exhaustive taxonomy

of every eligible work. Its common checkpoint is whether the resulting contrast and uncertainty still support the scientific or operational decision for which the experiment was run.

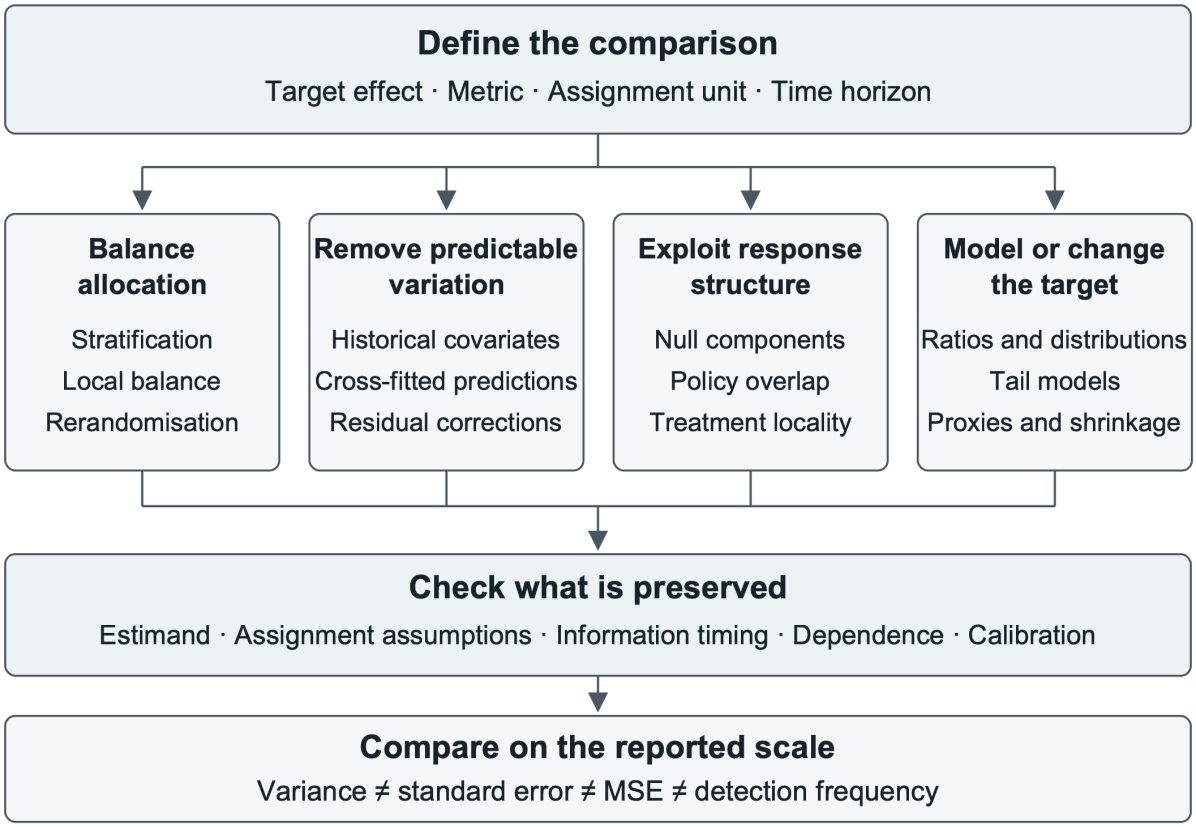

Figure 2. Mechanisms and validity checks for variance reduction in online controlled experiments. Categories overlap, and all require an explicit target, information set, and uncertainty calculation. The map is a conceptual synthesis of the reviewed methods.

5. Comparative evidence and validity

5.1 What can be compared

An alternative to assuming a known production effect is to hold out an independent experimental sample and use its difference in means as a noisy validation target. The meta-analysis of 699 Amazon supply-chain experiments uses this approach to compare estimators by relative mean-squared error. Under the stated independence and unbiasedness conditions, validation noise contributes a common term to that comparison; it does not disappear from an absolute error estimate. Product-level logistics interventions also differ from user-level engagement tests. This source therefore contributes an evaluation design and a domain-specific benchmark, without supplying a universal ranking for online metrics (Tripuraneni et al., 2023).

The strongest comparisons align the target, randomisation unit, information set, and baseline. Netflix's comparison of CUPED and poststratification is useful because it explicitly examines the use of the same covariates. In contrast, an apparent improvement obtained by giving one procedure a more informative predictor does not isolate an estimator effect. The same principle applies when historical data, auxiliary users, or additional outcomes are available to only one method. Those resources may be worth acquiring, but the

result should be attributed jointly to information and analysis rather than to an algorithmic name alone (Pei et al., 2026; Xie and Aurisset, 2016).

Even within a single study, different performance measures can rank methods differently. The ratio-metric evaluation reported different orderings by variance reduction and by sensitivity across its covariate variants. Its A/A Type-I error checks were near the nominal level, but those checks involved a particular experimental construction and did not prove general absence of bias. The sensitivity comparison used selected A/B pairs and one retention metric, which limits transfer to other outcomes. A review should retain these differences instead of selecting only the largest percentage from each source (Baweja et al., 2024).

Table 4 lists every quantitative precision or efficiency comparison explicitly recorded in the corpus coding - 22 entries from 16 of the 120 works - together with two recorded calibration and denominator results, organised by method family and retaining each original comparator and performance scale. Four families (network and cluster design; switchback and panel design; dynamic and dependence modelling; and sequential, multistage, and testing procedures) record no explicitly quantified precision comparison; that absence is part of the evidence profile. The boosted-tree variance ratio, the CUPED revenue result, the ReLOOP contrast count, and a projected power calculation have different denominators and evidential roles. No pooled variance-reduction percentage is calculated. An aggregate would require sufficiently compatible targets, assignment structures, baselines, uncertainty measures, and information about dependence between studies; that comparability is not established across this corpus, and the entries in Table 4 must not be averaged across rows (Deng et al., 2013; Hesterberg and Knight, 2024; Poyarkov et al., 2016; Sales et al., 2023).

Table 4. Quantitative precision and calibration comparisons recorded in the corpus coding, by method family, retaining the original performance scale and denominator.

Setting and source	Reported result	Comparator and scope	Interpretive limit
Pigeonhole covariate-cell design (Zhao and Zhou, 2025)	10.2% variance reduction and roughly 3% sample savings	Completely randomised assignment with the ordinary mean-difference estimator	Synthetic click outcomes on Yahoo visit covariates; one fixed realisation, 10,000 simulated assignments
Bing historical adjustment (Deng et al., 2013)	Revenue per user: less than 5% variance reduction; queries per user: about 50% with two-week history	Unadjusted versions of the respective metrics within one platform	Different metrics and correlations; not one transferable gain
Yandex boosted trees (Poyarkov et al., 2016)	Mean variance ratios: 0.3734 for trees, 0.3935 for all-feature linear adjustment, 0.4337 for one-feature linear adjustment	Unadjusted sessions metric set to 1; 161 experiments	Within-source mean of ratios; fitting and information sets must remain visible
Doubly robust ETL pipeline (Pasupuleti, 2023)	Standard error 0.95 (baseline), 0.60 (doubly robust), 0.56 (doubly robust with CUPED) in one representative synthetic throughput table	Baseline pipeline standard error in the same table	Synthetic scenarios; sample size, repetitions, outcome process, and sampling uncertainty are not reported
Bound on advanced versus basic adjustment (Ting and Hung, 2023)	Maximum additional variance reduction of 50% relative to basic single-covariate adjustment	Basic adjustment with one pre-period outcome covariate	Conditional analytical bound under a non-universal prediction-ordering assumption; no empirical validation

Continued on next page

Setting and source	Reported result	Comparator and scope	Interpretive limit
Marketplace replay evaluation (Staponaitė et al., 2025)	Confidence-interval width reductions: capping 36.7%, CUPED 21.2%, CUPAC 38.5%, doubly robust 21.1%	Unadjusted baseline in an injected-effect historical replay	Synthetic effects injected into control halves; interval coverage not reported
Ordered LOOP classroom trial (Pei et al., 2026)	Ordered LOOP standard error 0.0078 versus t-test 0.0083 in one real trial; standard LOOP 0.0033 versus 0.0030 in an adverse linear simulation	t-test standard error in the same trial or simulation	One real trial with unknown true effect; the simulation is an adverse case for the imputation model
Linearised ratio metrics at Yandex (Budylin et al., 2018)	Up to 34% additional detected tests	Per-user ratio metric without linearisation and boosted-tree adjustment	Detected-test count, not a variance reduction; selected experiment set from one platform
Empirical-Bayes arm shrinkage (Dimmery et al., 2019)	Arm-level mean-squared-error improvement stated as 85% and as around 90% at different points of the source	Unshrunk arm means	Data-calibrated simulations on 17 Facebook experiments; the two stated figures are inconsistent in the source
Cross-experiment shrinkage validation (Coey and Cunningham, 2019)	44% best predictive-error improvement	Pooled unshrunk effect estimates against a held-out validation estimate; 226 news-feed experiments	Predictive-error criterion; neither a pure variance reduction nor a new-experiment guarantee
Trimming, weighting, and CUPED combined (Liou and Taylor, 2020)	Average variance reduction 17.31% (variance-weighted estimation), 37.24% (CUPED), 48.38% (combined)	Standard unweighted, unadjusted estimator	Simulated noise plus 100 Facebook experiments; the components rest on different assumptions
Retention: historical controls (Baweja et al., 2024)	Variance change −72.47%; lower p-value in 15.38% of selected pairs	Pre-period numerator, denominator, and linearised metric; original one-day-retention ratio as baseline	13 selected A/B pairs; negative variance change means reduction
Retention: prediction controls (Baweja et al., 2024)	Variance change −45.66%; lower p-value in 76.92% of selected pairs	GBDT predictions of numerator, denominator, and linearised metric; same original ratio baseline	Same 13 selected pairs; lower-p-value frequency is not statistical power
Retention: combined controls (Baweja et al., 2024)	Variance change −72.62%; lower p-value in 30.77% of selected pairs	Both historical and predicted sets; same original ratio baseline	Same 13 selected pairs; variance and detection rankings differ
Learned proxy metric (Tripuraneni et al., 2024)	Held-out proxy quality 0.302 versus 0.279; sensitivity 0.181 versus 0.182 (essentially unchanged)	Baseline proxy on the same held-out test set	307 historical recommendation-system tests; a joint alignment-and-noise criterion, not a pure precision gain
Pareto-optimal proxy (Zito et al., 2025)	8.5-fold binary sensitivity relative to a short-term north-star metric	Short-term north-star metric on the same temporal hold-out	Retrospective construction; the proxy is not an unbiased estimator of the long-term effect
Campaign exposure filtering (Johnson et al., 2017b)	31% reported precision improvement (standard-error/confidence-interval reduction)	Exposure-scaled treatment-on-treated baseline in the same three-arm campaign	One 2010 apparel campaign; a standard-error reduction, not a variance percentage

Continued on next page

Setting and source	Reported result	Comparator and scope	Interpretive limit
One-sided triggering (Deng et al., 2023b)	Variance reductions: 99.2% and 98.3%; corresponding mean-zero diagnostic p-values: 10^{-34} and 0.485	Prediction and in-experiment balancing variants against the naive estimator in one company example	The largest gain fails its diagnostic; non-rejection for another variant does not prove zero bias
MLRATE simulation (Guo et al., 2021)	Coverage of nominal 95% intervals: 95.18% for boosted trees, 95.34% for elastic net, 94.88% unadjusted	10,000 repetitions of the specified nonlinear simulation	Calibration evidence for that generator, not a universal coverage guarantee
Educational-platform residualisation (Sales et al., 2023)	227 eligible pairwise contrasts from 68 A/B tests	Precision evaluation after source-specific exclusions from 383 contrasts and 84 tests	Contrasts can share arms; this is an evidence denominator, not 227 independent trials

The retention A/B comparisons lasted 7–14 days and involved 1.6–3.5 million users per group. A separate 220-pair A/A exercise assessed Type-I error; it is not the denominator of the sensitivity percentages in Table 4. The selected A/B pairs, shared metric, and single-source setting limit transportability of these comparisons (Baweja et al., 2024).

5.2 Conditional guarantees and implementation evidence

The contrast between guarantees is substantive. Residualised inverse-probability estimation establishes exact expectation under its independent-assignment construction (Gagnon-Bartsch et al., 2023). MLRATE compares limiting variance after its cross-fitting and regression steps (Guo et al., 2021). Efficient ratio estimation invokes nuisance-function and denominator conditions (Jin and Ba, 2023). These are answers to different inferential tasks. Together with the fixed-slope and fitted-slope distinction in Section 3, they motivate the reporting checks in Table 5.

Clustered and network experiments particularly expose the distinction between a point estimator and its standard error. Cluster regression adjustment has separate assumptions for the adjusted contrast and for the reported uncertainty, including a many-cluster approximation and restrictions on dependence. Training-loop interference can further break the interpretation of a conventional two-sample test even when empirical estimator variability falls. In another network augmentation, selectively observed interior nodes complicate the analogy with a randomly labelled sample. A familiar residual-plus-prediction form is consequently insufficient evidence of either unbiasedness or calibrated inference (Chen et al., 2026; Karrer et al., 2021; Si, 2024).

Assumptions about effect homogeneity deserve similar attention. Inverse-variance weighting is optimal within a stated common-mean model, but changes its interpretation when effects vary with the weighting variables. Efficient quantile-effect weighting under a common shift provides another example: the target restrictions are part of the gain. In the checked misspecification simulation, an efficient influence-function estimator had very poor coverage for the population mean difference, because that unrestricted target did not satisfy the model used for efficiency. This is a case-specific demonstration of a general comparison problem, not a universal assessment of the method (Athey et al., 2023; Liou and Taylor, 2020; Nie et al., 2020).

Table 5. Checks that connect an efficiency claim to an interpretable experimental estimate.

Check	Evidence to retain	Why the distinction matters
Target preservation	Treatment contrast, population, metric, and time horizon; record transformations explicitly	Proxy prediction, trimmed means, and common-shift effects can answer different estimands (Athey et al., 2023; Charette and Boudreault, 2025; Tripuraneni et al., 2024)
Information timing	Whether each feature precedes assignment; origin and timing of model training	Treatment-affected adjustment can change the effect; a balance-test non-rejection is diagnostic only (Deng et al., 2013; Lin and Crespo, 2026)
Training dependence	External training, cross-fitting, leave-one-out fitting, or reuse of the analysed outcomes	Theoretical guarantees apply to a specified fitting procedure, not to predictive performance alone (Guo et al., 2021; Poyarkov et al., 2016; Sales et al., 2023)
Assignment and dependence	Randomisation unit, clustering, interference, and temporal reuse	Independent observations cannot be assumed from a large event count; changing the unit can alter bias and variance (Hesterberg and Knight, 2024; Holtz and Aral, 2020; Zeng et al., 2026)
Guarantee and calibration	Separate unbiasedness, asymptotic variance, oracle bounds, empirical coverage, and diagnostics	A favourable result on one of these criteria does not establish all the others (Berman and Feit, 2024; Gagnon-Bartsch et al., 2023; Hirata et al., 2025; Jin and Ba, 2023)
Comparator and empirical independence	Covariates, baseline estimator, scale of improvement, reused arms, and reference effect	Unequal information sets and proxy ground truth can explain apparent performance differences (Masoero et al., 2023; Sales et al., 2023; Xie and Aurisset, 2016)

5.3 Diagnostics and the limits of observed evidence

A/A reassignment, empirical coverage, and balance tests answer different diagnostic tasks. A/A data can reveal faulty null calibration in the observed population, while a conditional-imbalance analysis can show how intervals behave under selected assignments. Neither supplies known non-null treatment effects for a production experiment. Likewise, disagreement with an unadjusted estimate measures discrepancy between estimators, not estimation bias against truth. The checked production-scale adjustment study uses a full-data linear estimate as a reference in part of its robustness analysis; that is a useful operational comparison but remains a proxy reference (Liou and Taylor, 2020; Masoero et al., 2023; Zhou et al., 2024).

Failure cases identify the limits of the available checks. The larger gain in one-sided triggering failed the method's mean-zero diagnostic (Deng et al., 2023b). In latent-stratum adjustment, a model check missed a harmful unequal-variance scenario (Berman and Feit, 2024). A multileaving counterexample refuted an earlier fidelity claim without invalidating its original simulated preference measurements (Oosterhuis and Rijke, 2017). The synthesis implication is that empirical success and formal validity are complementary evidence: a diagnostic needs power against the relevant failure, while a theorem needs assumptions appropriate to the deployment.

6. Discussion

The corpus supports the distinction introduced in Section 1 between three sources of precision: better experimental comparisons, better prediction of irrelevant outcome variation, and stronger restrictions on the response being estimated. This comparison describes mechanisms rather than ranking estimators. General balancing, network assignment, and temporal schedules change which comparisons the experiment creates. Covariate adjustment uses information within those comparisons. Tail models, treatment locality, and proxy

construction add assumptions about the outcome or target. Methods combine these sources, which explains why a single taxonomy label is insufficient for evaluating a deployed procedure.

For a conventional mean contrast, the general foundations sharpen the baseline against which specialised adjustment should be judged. Fixed controls admit an exact randomisation identity; fitted interacted regression provides a conditional asymptotic guarantee; flexible prediction requires an additional fitting argument. A method that beats an unadjusted difference may still offer little beyond a well-specified interacted baseline using the same predictors. The corpus partly meets this standard already: of the 25 primary covariate-adjustment works, 20 report a precision comparison against at least one covariate-adjusted baseline, four compare only against an unadjusted difference in means, and one reports no numeric comparison; the 25 works are identifiable by family in Appendix A. Netflix's same-covariate comparison makes this information-versus-estimator distinction concrete (Xie and Aurisset, 2016). Bounds on advanced versus basic adjustment likewise depend on the assumed ordering of predictive information, which may fail under different measurement periods or seasonality (Ting and Hung, 2023).

The more specialised families spend different scientific resources. Network designs require control of assignment and exposure definitions; switchback designs require assumptions about persistence; trigger methods require a defensible untreated component; and dynamic models require a transition or locality structure. These are costs of identification and implementation, not merely computational costs. A gain obtained from richer logs, a more restrictive target, or a stronger response model should be attributed to those choices. Holding only the algorithm name fixed obscures the basis of the improvement.

The evidence profile identifies a second distinction. Analytical and constructed evidence are widespread, but fewer than half of the main works report observed randomised contrasts. That does not discredit theoretical contributions: simulations provide known truth, and derivations clarify conditions that production comparisons cannot establish. Conversely, platform evidence tests operational relevance but often compares estimators against an unknown effect, reuses arms, or reports several correlated metrics. The review therefore interprets the evidence by what it identifies rather than placing every paper on a single hierarchy based on sample size or venue.

Publication coverage also differs across method families. Seven of the eleven primary switchback and panel-design works have no confirmed publication, and one further group has only an earlier published report; all nine metric-construction works have linked publication records. These counts locate where the synthesis depends on reports at different publication stages. They do not test whether conclusions would survive exclusion of unpublished material. The reader should draw a conclusion about stability rather than validity from this contrast. Switchback and panel methods are comparatively recent and circulate first as platform reports, whereas metric-construction work has had longer to complete peer review, so the asymmetry reflects field maturity and reporting channel. Its practical consequence is that conclusions about the switchback family rest more heavily on versions that may still change, which is why the first-report, inspected-version, and publication years introduced in Section 4.1 are recorded separately: they support version-specific interpretation of recent results while leaving methodological validity to the stated assumptions and evidence.

Prior reviews already address experimental sensitivity and practical design. The contribution here is the alignment of targets, information, assumptions, and evidence across the full included corpus. This facilitates more informative comparisons: common targets and covariates, uncertainty calibrated to assignment units, evaluation separated from training, and disclosure of repeated use of the same experiments. The available data do not support an average cross-method percentage reduction or a universally best estimator. A future comparative benchmark would need to fix those design choices before evaluating algorithms, rather than aggregate the largest reported gains.

Several limitations remain in the evidence itself. Proprietary platform data constrain independent reanalysis.

Multiple reports, metrics, and experimental arms can reflect overlapping observations despite work-level deduplication. Source claims of general efficiency may refer to an oracle, an approximation, or one component of uncertainty. Null reassignment does not supply evidence about every non-null intervention, and proxy ground truth can distort comparisons. Publication status alone does not establish empirical independence or validity.

The review's process limitations, noted at their points of origin in Sections 2.1 and 2.3, are collected here. Two title-and-abstract/repository routes cannot establish complete coverage of a dispersed literature, especially where metadata are incomplete. English-language and full-text-access requirements can select which evidence is visible. The predecessor entitled "Efficient experimentation: A review of four methods to reduce the costs of A/B testing" was unavailable in full, limiting comparison with its coverage; none is inferred from its title. The single citation pass is deliberately bounded, and exact-item metadata checks do not establish search saturation. Access restrictions affected follow-up clarification, while the principal eight harvests were completed.

The lack of preregistration and retrospective coding limit the ability to distinguish prespecified classifications from choices informed by the retrieved literature.

7. Conclusion

The 120 main works describe a broad collection of ways to use experimental structure and auxiliary information more efficiently. Covariate adjustment is prominent, but design, network and temporal dependence, difficult outcomes, trigger information, metric construction, and sequential decisions contribute distinct mechanisms. Their effects on precision depend on the estimand and the information or assumptions used to obtain the gain.

The decisive comparison is between fully specified procedures. Exact fixed-coefficient identities, fitted-regression asymptotics, prediction-assisted estimators, and model-restricted efficiency each establish different results. Reported variance, mean-squared error, coverage, and detection frequency should remain separate, with comparator and empirical denominator attached. The practical implication is to specify the target and assignment first, identify the available unaffected information and modelling restrictions, and then compare precision together with bias and calibrated uncertainty.

Declarations

Ethics and research integrity. Not applicable. This is a systematic review of published literature and involved no human participants, animals or personal data.

Funding statement. This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Data Availability Statement. No new data were created. All works analysed are sources listed in Appendix A or cited in the References. The included-work list, descriptive coding, appraisal outcomes, and query specifications are reported in Appendices A and B; underlying extraction records and the version-grouping map are available from the author on reasonable request.

Use of AI writing tools. Computational agents assisted with screening, extraction and descriptive coding, as described in Section 2. All judgements were checked by the author, who takes full responsibility for the content. No AI tool is an author.

Conflicts of interest. The author declares no conflicts of interest.

References

1. Arbour, D., Ben-Michael, E., Feller, A., Lal, A., and Yuan, L.-H. (2026). AI-Assisted variance reduction in randomized experiments. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining v.2*, pp. 21–32. https://doi.org/10.1145/3770855.3817896.

2. Athey, S., Bickel, P. J., Chen, A., Imbens, G. W., and Pollmann, M. (2023). Semi-parametric estimation of treatment effects in randomised experiments. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **85**(5), 1615–1638. https://doi.org/10.1093/jrsssb/qkad072.

3. Auer, F., Ros, R., Kaltenbrunner, L., Runeson, P., and Felderer, M. (2021). Controlled experimentation in continuous experimentation: Knowledge and challenges. *Information and Software Technology*, **134**, 106551. https://doi.org/10.1016/j.infsof.2021.106551.

4. Bajari, P., Burdick, B., Imbens, G. W., Masoero, L., McQueen, J., Richardson, T. S., and Rosen, I. M. (2023). Experimental design in marketplaces. *Statistical Science*, **38**(3). https://doi.org/10.1214/23-sts883.

5. Baweja, S., Pokharna, N., Ustimenko, A., and Jeunen, O. (2024). Variance reduction in ratio metrics for efficient online experiments. In *Advances in Information Retrieval*, pp. 292–297. https://doi.org/10.1007/978-3-031-56069-9_34.

6. Berman, R. and Feit, E. M. (2024). Latent stratification for incrementality experiments. *Marketing Science*, **43**(4), 903–917. https://doi.org/10.1287/mksc.2022.0297.

7. Bojinov, I., Simchi-Levi, D., and Zhao, J. (2023). Design and analysis of switchback experiments. *Management Science*, **69**(7), 3759–3777. https://doi.org/10.1287/mnsc.2022.4583.

8. Brost, B., Cox, I. J., Seldin, Y., and Lioma, C. (2016). An improved multileaving algorithm for online ranker evaluation. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 745–748. https://doi.org/10.1145/2911451.2914706.

9. Budylin, R., Drutsa, A., Katsev, I., and Tsoy, V. (2018). Consistent transformation of ratio metrics for efficient online controlled experiments. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pp. 55–63. https://doi.org/10.1145/3159652.3159699.

10. Chapelle, O., Joachims, T., Radlinski, F., and Yue, Y. (2012). Large-scale validation and analysis of interleaved search evaluation. *ACM Transactions on Information Systems*, **30**(1), 1–41. https://doi.org/10.1145/2094072.2094078.

11. Charette, K. and Boudreault, T. (2025). Improving sensitivity in A/B tests: Integrating CUPED with trimmed mean techniques. arXiv preprint 2510.03468v1. https://doi.org/10.48550/arxiv.2510.03468.

12. Chen, Q., Wu, A., Li, B., Deng, L., and Wang, Y. (2026). Journey to the centre of cluster: Harnessing interior nodes for A/B testing under network interference. In *International Conference on Learning Representations*, pp. 20461–20480. https://proceedings.iclr.cc/paper_files/paper/2026/hash/22a25fc3da528794d52664dacc7bd470-Abstract-Conference.html.

13. Chen, S., Simchi-Levi, D., and Wang, C. (2025). Improving the estimation of lifetime effects in A/B testing via treatment locality. arXiv preprint 2407.19618v3. https://doi.org/10.48550/arxiv.2407.19618.

14. Chihara, N., Oka, T., Matsubara, Y., Sakurai, Y., and Yasui, S. (2026). Modeling covariate transition for efficient estimation of longitudinal treatment effects in randomized experiments. arXiv preprint 2605.31443v1. https://arxiv.org/abs/2605.31443v1.

15. Coey, D. and Cunningham, T. (2019). Improving treatment effect estimators through experiment splitting. In *The World Wide Web Conference*. https://doi.org/10.1145/3308558.3313452.

16. Deng, A. (2015). Objective Bayesian two sample hypothesis testing for online controlled experiments. In *Proceedings of the 24th International Conference on World Wide Web*, pp. 923–928. https://doi.org/10.1145/2740908.2742563.

17. Deng, A., Du, M., Matlin, A., and Zhang, Q. (2023a). Variance reduction using in-experiment data: Efficient and targeted online measurement for sparse and delayed outcomes. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3937–3946. https://doi.org/10.1145/3580305.3599928.

18. Deng, A., Hagar, L., Stevens, N. T., Xifara, T., and Gandhi, A. (2024). Metric decomposition in A/B tests. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4885–4895. https://doi.org/10.1145/3637528.3671556.

19. Deng, A. and Hu, V. (2015). Diluted treatment effect estimation for trigger analysis in online controlled experiments. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pp. 349–358. https://doi.org/10.1145/2684822.2685307.

20. Deng, A., Li, T., and Guo, Y. (2014). Statistical inference in two-stage online controlled experiments with treatment selection and validation. In *Proceedings of the 23rd International Conference on World Wide Web*, pp. 609–618. https://doi.org/10.1145/2566486.2568028.

21. Deng, A., Xu, Y., Kohavi, R., and Walker, T. (2013). Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, pp. 123–132. https://doi.org/10.1145/2433396.2433413.

22. Deng, A., Yuan, L.-H., Kanai, N., and Salama-Manteau, A. (2023b). Zero to hero: Exploiting null effects to achieve variance reduction in experiments with one-sided triggering. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pp. 823–831. https://doi.org/10.1145/3539597.3570413.

23. Dimmery, D., Bakshy, E., and Sekhon, J. (2019). Shrinkage estimators in online experiments. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2914–2922. https://doi.org/10.1145/3292500.3330771.

24. Drutsa, A., Gusev, G., and Serdyukov, P. (2015). Future user engagement prediction and its application to improve the sensitivity of online experiments. In *Proceedings of the 24th International Conference on World Wide Web*, pp. 256–266. https://doi.org/10.1145/2736277.2741116.

25. Fithian, W. and Ting, D. (2017). Family learning: Non-parametric statistical inference with parametric efficiency. arXiv preprint 1711.10028v1. https://arxiv.org/abs/1711.10028v1.

26. Fithian, W. and Wager, S. (2015). Semiparametric exponential families for heavy-tailed data. *Biometrika*, **102**(2), 486–493. https://doi.org/10.1093/biomet/asu065.

27. Freedman, D. A. (2008a). On regression adjustments in experiments with several treatments. *The Annals of Applied Statistics*, **2**(1), 176–196. https://doi.org/10.1214/07-aoas143.

28. Freedman, D. A. (2008b). On regression adjustments to experimental data. *Advances in Applied Mathematics*, **40**(2), 180–193. https://doi.org/10.1016/j.aam.2006.12.003.

29. Gagnon-Bartsch, J. A., Sales, A. C., Wu, E., Botelho, A. F., Erickson, J. A., Miratrix, L. W., and Heffernan, N. T. (2023). Precise unbiased estimation in randomized experiments using auxiliary observational data. *Journal of Causal Inference*, **11**(1), 20220011. https://doi.org/10.1515/jci-2022-0011.

30. Guo, Y., Coey, D., Konutgan, M., Li, W., Schoener, C., and Goldman, M. (2021). Machine learning for variance reduction in online experiments. In *Advances in Neural Information Processing Systems 34*, pp. 8637–8648. https://proceedings.neurips.cc/paper/2021/hash/488b084119a1c7a4950f00706ec7ea16-Abstract.html.

31. Ham, D. W., Lindon, M., Tingley, M., and Bojinov, I. (2026). Design-based confidence sequences: A general approach to risk mitigation in panel experiments. *Management Science*, mnsc.2024.08235. https://doi.org/10.1287/mnsc.2024.08235.

32. Hesterberg, T. and Knight, B. (2024). Power analysis for experiments with clustered data, ratio metrics, and regression for covariate adjustment. arXiv preprint 2406.06834v1. https://doi.org/10.48550/arxiv.2406.06834.

33. Hirata, T., Byambadalai, U., Oka, T., Yasui, S., and Uto, S. (2025). Efficient and scalable estimation of distributional treatment effects with multi-task neural networks. arXiv preprint 2507.07738v1. https://arxiv.org/abs/2507.07738v1.

34. Holtz, D. and Aral, S. (2020). Limiting bias from test-control interference in online marketplace experiments. arXiv preprint 2004.12162v1. https://arxiv.org/abs/2004.12162v1.

35. Hu, Y. and Wager, S. (2026). Switchback experiments under geometric mixing. arXiv preprint 2209.00197v5. https://arxiv.org/abs/2209.00197v5.

36. Iizuka, K., Seki, Y., and Kato, M. P. (2021). Decomposition and interleaving for variance reduction of post-click metrics. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*, pp. 221–230. https://doi.org/10.1145/3471158.3472235.

37. Jeunen, O. (2026). Accelerating A/B-Tests with counterfactual estimation: Reducing variance through policy overlap. arXiv preprint 2607.14604v1. https://doi.org/10.48550/arxiv.2607.14604.

38. Jeunen, O. and Ustimenko, A. (2024). Learning metrics that maximise power for accelerated A/B-Tests. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5183–5193. https://doi.org/10.1145/3637528.3671512.

39. Jin, Y. and Ba, S. (2023). Toward optimal variance reduction in online controlled experiments. *Technometrics*, **65**(2), 231–242. https://doi.org/10.1080/00401706.2022.2142670.

40. Johnson, G. A., Lewis, R. A., and Nubbemeyer, E. I. (2017a). Ghost ads: Improving the economics of measuring online ad effectiveness. *Journal of Marketing Research*, **54**(6), 867–884. https://doi.org/10.1509/jmr.15.0297.

41. Johnson, G. A., Lewis, R. A., and Reiley, D. H. (2017b). When less is more: Data and power in advertising experiments. *Marketing Science*, **36**(1), 43–53. https://doi.org/10.1287/mksc.2016.0998.

42. Karrer, B., Shi, L., Bhole, M., Goldman, M., Palmer, T., Gelman, C., Konutgan, M., and Sun, F. (2021). Network experimentation at scale. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 3106–3116. https://doi.org/10.1145/3447548.3467091.

43. Kohavi, R., Longbotham, R., Sommerfield, D., and Henne, R. M. (2009). Controlled experiments on the web: Survey and practical guide. *Data Mining and Knowledge Discovery*, **18**(1), 140–181. https://doi.org/10.1007/s10618-008-0114-1.

44. Larsen, N., Stallrich, J., Sengupta, S., Deng, A., Kohavi, R., and Stevens, N. T. (2024). Statistical challenges in online controlled experiments: A review of A/B testing methodology. *The American Statistician*, **78**(2), 135–149. https://doi.org/10.1080/00031305.2023.2257237.

45. Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *The Annals of Applied Statistics*, **7**(1), 295–318. https://doi.org/10.1214/12-aoas583.

46. Lin, Z. and Crespo, P. (2026). Variance reduction combining pre-experiment and in-experiment data. In *Proceedings of Machine Learning Research 323*, pp. 699–717. https://proceedings.mlr.press/v323/lin26a.html.

47. Liou, K. and Taylor, S. J. (2020). Variance-weighted estimators to improve sensitivity in online experiments. In *Proceedings of the 21st ACM Conference on Economics and Computation*, pp. 837–850. https://doi.org/10.1145/3391403.3399542.

48. Liu, M., Sun, X., Varshney, M., and Xu, Y. (2018). Large-scale online experimentation with quantile metrics. In *JSM 2018 Proceedings, Section on Statistical Consulting*, pp. 2849–2860. https://ww2.amstat.org/meetings/proceedings/2018/data/assets/pdf/867117.pdf.

49. Lopes, H. F., Taddy, M., and Gardner, M. (2019). Scalable semiparametric inference for the means of heavy-tailed distributions. In *Topics in Identification, Limited Dependent Variables, Partial Observability, Experimentation, and Flexible Modeling: Part b*, pp. 141–156. https://doi.org/10.1108/s0731-90532019000040b008.

50. Masoero, L., Hains, D., and McQueen, J. (2023). Leveraging covariate adjustments at scale in online A/B testing. In *Proceedings of the KDD'23 Workshop on Causal Discovery, Prediction and Decision*, pp. 25–48. https://proceedings.mlr.press/v218/masoero23a.html.

51. Nie, K., Kong, Y., Yuan, T. T., and Burke, P. B. (2020). Dealing with ratio metrics in A/B testing at the presence of intra-user correlation and segments. In *Web Information Systems Engineering – WISE 2020*, pp. 563–577. https://doi.org/10.1007/978-3-030-62008-0_39.

52. Oosterhuis, H. and Rijke, M. de (2017). Sensitive and scalable online evaluation with theoretical guarantees. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pp. 77–86. https://doi.org/10.1145/3132847.3132895.

53. Oosterhuis, H. and Rijke, M. de (2020). Taking the counterfactual online: Efficient and unbiased online evaluation for ranking. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*, pp. 137–144. https://doi.org/10.1145/3409256.3409820.

54. Pasupuleti, M. K. (2023). GenAI-assisted ETL on Hadoop: Petascale batch causal A/B pipelines. *International Journal of Academic and Industrial Research Innovations*. https://doi.org/10.62311/nesx/rp-30032023-41-52.

55. Pei, Y., Sales, A., Gagnon-Bartsch, J., Worden, E., and Heffernan, N. (2026). Accounting for classroom effects in student-level randomized online experiments: The ordered LOOP estimator. In *Proceedings of the 19th International Conference on Educational Data Mining*, pp. 589–593. https://doi.org/10.5281/zenodo.21039790.

56. Pokharna, N., Jeunen, O., Saraf, Y., and Ustimenko, A. (2026). Variance reduction for heavy-tailed monetization metrics in ranking experiments via post-stratification. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 4832–4836. https://doi.org/10.1145/3805712.3808428.

57. Poyarkov, A., Drutsa, A., Khalyavin, A., Gusev, G., and Serdyukov, P. (2016). Boosted decision tree regression adjustment for variance reduction in online controlled experiments. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 235–244. https://doi.org/10.1145/2939672.2939688.

58. Quin, F., Weyns, D., Galster, M., and Silva, C. C. (2023). A/B testing: A systematic literature review. arXiv preprint 2308.04929v1. https://doi.org/10.48550/arxiv.2308.04929.

59. Radlinski, F. and Craswell, N. (2013). Optimized interleaving for online retrieval evaluation. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, pp. 245–254. https://doi.org/10.1145/2433396.2433429.

60. Sakhi, O., Gilotte, A., and Rohde, D. (2026). Exploiting similarities in A/B testing with off-policy estimation. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining v.2*, pp. 4103–4114. https://doi.org/10.1145/3770855.3817788.

61. Sales, A. C., Prihar, E. B., Gagnon-Bartsch, J. A., and Heffernan, N. T. (2023). Using auxiliary data to boost precision in the analysis of A/B tests on an online educational platform: New data and new results. *Journal of Educational Data Mining*, **15**(2), 53–85. https://doi.org/10.5281/zenodo.8016854.

62. Si, N. (2024). Tackling interference induced by data training loops in A/B tests: A weighted training approach. arXiv preprint 2310.17496v5. https://doi.org/10.48550/arxiv.2310.17496.

63. Staponaitė, G., Gamper, J., Giniūnaitė, R., and Reklaitė, A. (2025). Variance reduction in online marketplace A/B testing. In *KDD 2025 Undergraduate and Masters Consortium*.

64. Ting, D. and Hung, K. (2023). On the limits of regression adjustment. arXiv preprint 2311.17858v1. https://arxiv.org/abs/2311.17858v1.

65. Tripuraneni, N., Perrault-Joncas, D., Madeka, D., Foster, D., and Jordan, M. I. (2023). Meta-analysis of randomized experiments with applications to heavy-tailed response data. arXiv preprint 2112.07602v5. https://arxiv.org/abs/2112.07602v5.

66. Tripuraneni, N., Richardson, L., D'Amour, A., Soriano, J., and Yadlowsky, S. (2024). Choosing a proxy metric from past experiments. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5803–5812. https://doi.org/10.1145/3637528.3671543.

67. Wang, W. and Zhang, X. (2021). CONQ: CONtinuous quantile treatment effects for large-scale online controlled experiments. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pp. 202–210. https://doi.org/10.1145/3437963.3441779.

68. Wu, Y., Zheng, Z., Zhang, G., Zhang, Z., and Wang, C. (2022). Non-stationary A/B tests. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2079–2089. https://doi.org/10.1145/3534678.3539325.

69. Xie, H. and Aurisset, J. (2016). Improving the sensitivity of online controlled experiments. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 645–654. https://doi.org/10.1145/2939672.2939733.

70. Xiong, R., Chin, A., and Taylor, S. J. (2024). Data-driven switchback experiments: Theoretical tradeoffs and empirical bayes designs. arXiv preprint 2406.06768v1. https://arxiv.org/abs/2406.06768v1.

71. Zeng, Z., Adjaho, C., Bucarey, A., Qin, C., Zhang, R., Hoban, P., Johari, R., and Wager, S. (2026). Sequentially-rerandomized switchback experiments. arXiv preprint 2604.02489v1. https://arxiv.org/abs/2604.02489v1.

72. Zhang, Q. (2025). Rerandomization algorithms for optimal designs of network A/B tests. *Technometrics*, **67**(4), 655–668. https://doi.org/10.1080/00401706.2025.2505438.

73. Zhang, Q., Deng, A., Du, M., Gao, H., He, L., and Katariya, S. (2025a). Harnessing the power of interleaving and counterfactual evaluation for Airbnb search ranking. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining v.2*, pp. 5205–5214. https://doi.org/10.1145/3711896.3737232.

74. Zhang, Y., Wan, B., and Qin, Y. (2025b). Bridging control variates and regression adjustment in A/B testing: From design-based to model-based frameworks. arXiv preprint 2509.13944v2. https://doi.org/10.48550/arxiv.2509.13944.

75. Zhao, J. and Zhou, Z. (2025). Pigeonhole design: Balancing sequential experiments from an online matching perspective. *Management Science*, **71**(3), 1889–1908. https://doi.org/10.1287/mnsc.2023.02184.

76. Zhou, H., Sun, K., Li, S., Fan, Y., Jiang, G., Zheng, J., and Li, T. (2024). STATE: A robust ATE estimator of heavy-tailed metrics for variance reduction in online controlled experiments. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 6380–6389. https://doi.org/10.1145/3637528.3672352.

77. Zito, A., Greaves, D., Soriano, J., and Richardson, L. (2025). Pareto optimal proxy metrics. *Applied Stochastic Models in Business and Industry*, **41**(2), e70003. https://doi.org/10.1002/asmb.70003.

Appendix A. Included works, descriptive coding, and appraisal

Table A1 lists all 120 main works. Family codes: F1 general allocation, balance, and stratification; F2 network and cluster design; F3 switchback and panel design; F4 covariate, control-variate, and prediction adjustment; F5 ratio, robust, distributional, and shrinkage estimation; F6 metric construction and proxies; F7 trigger, exposure, and policy information; F8 dynamic and dependence modelling; F9 sequential, multistage, and testing procedures; secondary families follow a plus sign. Evidence codes: A analytical argument or derivation; S simulation or constructed experiment; O observed randomised contrasts; E other empirical data. Publication status at the 25 September 2026 cutoff: P publication record linked to the work; E earlier report published only; R repository or report without confirmed publication; U unresolved. The appraisal column gives the eight domain judgements in the order of Table 2 (target/baseline, validity assumptions,

training/evaluation separation, simulation description, empirical provenance, Type-I error and coverage, code and data availability, transportability), coded A adequate, C concern, U unclear, – not applicable. IDs are the stable selection identifiers of the review records.

Table A1. The 120 main works.

ID	Work	Family	Ev.	Pub	Appraisal
C-BC004	Koki Konishi (2026). A More Accurate Algorithm Comparison through A/B Testing using Of…	F7	ASE	P	AC–CCUAC
C-BC005	Otmane Sakhi (2026). Exploiting Similarities in A/B Testing with Off-Policy Estimation	F7	AS	P	ACUC–UUC
C-BC006	Harrie Oosterhuis (2020). Taking the Counterfactual Online: Efficient and Unbiased Online E…	F7 (+F1)	ASE	P	ACCCCUAC
C-BC007	Qing Zhang (2025). Harnessing the Power of Interleaving and Counterfactual Evaluatio…	F7 (+F6)	AO	P	CCU–CUUC
C-BC008	Olivier Jeunen (2024). Learning Metrics that Maximise Power for Accelerated A/B-Tests	F6	AO	P	CCC–CACC
C-BC009	Olivier Chapelle (2012). Large-Scale Validation and Analysis of Interleaved Search Evaluation	F7 (+F6)	AO	P	CCA–CUUC
C-BC017	Ozan Candogan (2023). Correlated Cluster-Based Randomized Experiments: Robust Variance…	F2	ASE	P	AC–CCCUC
C-BC018	Huan Gui (2015). Network A/B Testing: From Sampling to Estimation	F2 (+F8)	ASOE	P	AC–CCUUC
C-BC019	Yang Liu (2022). Adaptive A/B Test on Networks with Cluster Structures	F2	ASE	P	AC–CCCAC
C-BC020	Victoria Pokhiko (2019). D-optimal Design for Network A/B Testing	F2	ASE	P	AC–CCUUC
C-BC021	Qiong Zhang (2022). Locally Optimal Design for A/B Tests in the Presence of Covariate…	F2 (+F1)	ASE	P	AC–CCUUC
C-BC022	Johan Ugander (2013). Graph Cluster Randomization: Network Exposure to Multiple Universes	F2 (+F7)	AS	P	AC–C–UUC
C-BC025	Avery Closser (2024). Should we account for classrooms? Analyzing online experimental d…	F4	ASE	P	ACACCUUC
C-BC026	Adam C Sales (2018). Using Big Data to Sharpen Design-Based Inference in A/B Tests	F4	AOE	P	ACA–CUUC
C-BC028	Iavor Bojinov (2025). Design and Analysis of Switchback Experiments	F3	AS	P	AC–C–CUC
C-BC029	Vivek F. Farias (2022). Markovian Interference in Experiments	F8 (+F7)	ASE	P	ACCC–CAC
C-BC030	Vivek F. Farias (2023). Correcting for Interference in Experiments: A Case Study at Douyin	F8 (+F4)	ASE	P	ACCCCCUC
C-BC031	Peter Glynn (2020). Adaptive Experimental Design with Temporal Interference: A Maximu…	F3 (+F8)	A	P	AC——AUC
C-BC033	Yuchen Hu (2026). Switchback Experiments under Geometric Mixing	F3 (+F8)	ASE	R	ACCC–CUC
C-BC034	Hannah Li (2021). Interference, Bias, and Variance in Two-Sided Marketplace Experim…	F2 (+F1)	AS	P	AC–C–UUC
C-BC035	Shuangning Li (2026). Experimenting Under Stochastic Congestion	F8 (+F3)	ASE	R	AC–A–CUC
C-BC037	Jean Pouget-Abadie (2019). Variance Reduction in Bipartite Experiments through Correlation C…	F2	ASE	P	CC–C–UUC

Continued on next page

ID	Work	Family	Ev.	Pub	Appraisal
C-BC038	Johan Ugander (2020). Randomized Graph Cluster Randomization	F2 (+F7)	ASE	P	AC–A–UUC
C-BC041	Daniel Ting (2023). On the Limits of Regression Adjustment	F4	A	R	AA———C
C-BC042	Nilesh Tripuraneni (2023). Choosing a Proxy Metric from Past Experiments	F6	AO	P	AAAACUUC
C-BC046	Brian Karrer (2020). Network experimentation at scale	F4 (+F2+F7)	ASO	P	AACACUUC
C-BC048	William Fithian (2017). Family Learning: Non-Parametric Statistical Inference with Parame…	F5 (+F6)	ASO	R	ACAACAUC
C-BC049	Roman Budylin (2018). Consistent Transformation of Ratio Metrics for Efficient Online C…	F5 (+F4+F6)	ASO	P	ACUACAUC
C-BC060	William Fithian (2013). Semiparametric Exponential Families for Heavy-Tailed Data	F5	ASE	P	ACCACUUC
C-BC061	Matt Taddy (2014). A nonparametric Bayesian analysis of heterogeneous treatment effe…	F5 (+F4+F1)	AO	P	ACU–CUUC
C-BC064	Shikai Luo (2022). Policy Evaluation for Temporal and/or Spatial Dependent Experimen…	F8 (+F3)	ASO	P	ACCAAACC
C-BC065	Chengchun Shi (2022). Dynamic Causal Effects Evaluation in A/B Testing with a Reinforce…	F8 (+F9+F3)	ASO	P	CCCACAAC
C-BC066	Chengchun Shi (2023). A Multi-Agent Reinforcement Learning Framework for Off-Policy Eva…	F8 (+F7)	ASE	P	ACCACUAC
C-BC067	Yifan Zhou (2020). Cluster-Adaptive Network A/B Testing: From Randomization to Estim…	F2 (+F1+F7)	ASE	P	CC–A–UUC
C-BC072	Nilesh Tripuraneni (2021). Meta-Analysis of Randomized Experiments with Applications to Heav…	F5 (+F4)	AO	U	ACA–CUUC
C-BC074	Aiyou Chen (2021). Robust Causal Inference for Incremental Return on Ad Spend with R…	F5 (+F1)	ASO	P	ACCAACAC
C-BC085	Sergei Pankratev (2026). Powerful Switchback Experiments – Or Not?	F3 (+F4)	AS	R	CC–C–UUC
C-BC086	Ruoxuan Xiong (2024). Data-Driven Switchback Experiments: Theoretical Tradeoffs and Emp…	F3	ASE	E	ACCACCUC
C-BC092	Reza Hosseini (2019). Unbiased variance reduction in randomized experiments	F4 (+F5)	AS	R	ACCC–UUC
C-BC096	Xinrui Ruan (2026). Can Language Models Boost the Power of Randomized Experiments Wit…	F4	ASO	R	ACACAAAC
C-BC103	Olivier Jeunen (2024). Powerful A/B-Testing Metrics and Where to Find Them	F6 (+F9)	AO	P	ACUUCCUC
C-BC127	Garrett A. Johnson (2015). When Less is More: Data and Power in Advertising Experiments	F7 (+F4)	AO	P	ACC–CUUC
C-BC128	Garrett A. Johnson (2015). Ghost Ads: Improving the Economics of Measuring Ad Effectiveness	F7	AO	P	ACA–CUUC
C-BC130	Dean Eckles (2014). Design and Analysis of Experiments in Networks: Reducing Bias fro…	F2 (+F7)	AS	P	AC–A–UUC
C-BC134	Martin Saveski (2017). Detecting Network Effects: Randomizing Over Randomized Experiments	F2 (+F1+F4)	ASO	P	ACA–ACUC
C-BC136	Paul Missault (2025). Robust and efficient multiple-unit switchback experimentation	F3	ASOE	R	AC–CCAUC

Continued on next page

ID	Work	Family	Ev.	Pub	Appraisal
C-BC138	Tu Ni (2025). Reliable Switchback Experiments with Rerandomization for Auction…	F3 (+F1)	AS	R	ACAACCUC
C-BC142	Su Jia (2025). Clustered Switchback Designs for Experimentation Under Spatio-tem…	F3 (+F2+F7)	AS	R	AC–A–UUC
C-BC146	Alex Deng (2014). Statistical Inference in Two-Stage Online Controlled Experiments…	F9 (+F5)	AS	P	ACAA–AUC
C-BC151	Alex Deng (2015). Objective Bayesian Two Sample Hypothesis Testing for Online Contr…	F9 (+F5)	ASO	P	AACACUUC
C-BC155	Qianyi Chen (2023). Optimized Covariance Design for AB Test on Social Network under I…	F2	ASE	P	ACCA–UAC
C-BC156	Qianyi Chen (2025). Just Ramp-up: Unleash the Potential of Regression-based Estimator…	F8 (+F2+F9)	ASE	R	ACCA–UUC
C-BC157	Shaojie Deng (2011). Choice of the Randomization Unit in Online Controlled Experiment	F1	AS	U	AC–ACUUC
C-BC158	Jason (Xiao) Wang (2019). Measuring Average Treatment Effect from Heavy-tailed Data	F5 (+F4)	ASE	R	CCACACUC
C-BC159	Sergei Pankratev (2026). Power-Optimal Covariate Adjustment for Switchback Experiments	F4	AS	R	ACUA–CUC
C-BC162	Alex Deng (2016). Data-Driven Metric Development for Online Controlled Experiments:…	F6 (+F5)	AO	P	AAC–CCUC
C-BC165	Tu Ni (2024). Design of Panel Experiments with Spatial and Temporal Interference	F3 (+F2+F7)	ASE	R	AC–AAAUC
C-BC167	Ruoxuan Xiong (2023). Optimal Experimental Design for Staggered Rollouts	F3 (+F1+F4)	ASE	P	ACAACAAC
C-BC169	Nikhil Bhat (2017). Near Optimal A-B Testing	F1	ASE	P	ACAACUUC
C-BC172	Pavel Dmitriev (2016). Measuring Metrics	F6	AO	P	ACC–CUCC
C-BC187	Eugene Kharitonov (2017). Learning Sensitive Combinations of A/B Test Metrics	F6	ASO	P	CCA–ACUC
C-BC194	Anne Schuth (2015). Predicting Search Satisfaction Metrics with Interleaved Comparisons	F7 (+F6)	AO	P	ACA–CUUC
S-0001	Alex Deng (2013). Improving the Sensitivity of Online Controlled Experiments by Uti…	F4 (+F1+F5)	AO	P	AAC–UUUC
S-0002	Huizhi Xie (2016). Improving the Sensitivity of Online Controlled Experiments: Case…	F1 (+F4)	ASE	P	AAAAUUUC
S-0003	Alexey Poyarkov (2016). Boosted Decision Tree Regression Adjustment for Variance Reductio…	F4	ASO	P	ACCAUACC
S-0004	Ying Jin (2023). Toward Optimal Variance Reduction in Online Controlled Experiments	F4 (+F5)	ASO	P	AAAAUAUC
S-0005	Yongyi Guo (2021). Machine Learning for Variance Reduction in Online Experiments	F4	ASE	P	AAAAUAUC
S-0006	Shubham Baweja (2024). Variance Reduction in Ratio Metrics for Efficient Online Experiments	F5 (+F4)	AO	P	ACU–CAUC
S-0007	Lorenzo Masoero (2023). Leveraging covariate adjustments at scale in online A/B testing	F4	ASO	P	ACUAUAUC
S-0009	Drew Dimmery (2019). Shrinkage Estimators in Online Experiments	F5 (+F9)	ASO	P	AA–ACCUC

Continued on next page

ID	Work	Family	Ev.	Pub	Appraisal
S-0010	Alex Deng (2024). Metric Decomposition in A/B Tests	F6 (+F7+F5)	ASO	P	ACUAUAAC
S-0011	Yuhang Wu (2022). Non-stationary A/B Tests	F1 (+F3)	ASE	P	AA–ACUUC
S-0016	Kevin Liou (2020). Variance-Weighted Estimators to Improve Sensitivity in Online Exp…	F5 (+F4)	ASO	P	CCUAUUUC
S-0017	Shuze Chen (2025). Improving the Estimation of Lifetime Effects in A/B Testing via T…	F8 (+F7)	AS	R	AA–ACUUC
S-0018	Kojiro Iizuka (2021). Decomposition and Interleaving for Variance Reduction of Post-cli…	F7 (+F6)	ASE	P	ACUACUAC
S-0019	Alex Deng (2023). Zero to Hero: Exploiting Null Effects to Achieve Variance Reducti…	F7 (+F4)	ASO	P	AAAAUUAC
S-0025	Hao Zhou (2024). STATE: A Robust ATE Estimator of Heavy-Tailed Metrics for Varianc…	F5 (+F4)	ASE	P	CCAACUUC
S-0027	Haru Momozu (2025). Subset Selection for Stratified Sampling in Online Controlled Exp…	F1	ASE	P	ACAACUUC
S-0028	Tomoka Takei (2026). Optimal Multi-way Decision Trees for Stratified Sampling in Onlin…	F1	ASE	U	AAAACUAC
S-0029	Yu Zhang (2025). Bridging Control Variates and Regression Adjustment in A/B Testin…	F4	ASE	R	AAUACAUC
S-0030	Kevin Charette (2025). Improving Sensitivity in A/B Tests: Integrating CUPED with Trimme…	F5 (+F4+F6)	AS	R	ACUA–AUC
S-0031	Sergei Pankratev (2026). CUPED on Steroids: Multivariate Covariate Adjustment for Switchba…	F4	ASE	R	CUUACUUC
S-0032	Yu Zhang (2026). Ensuring Trustworthy Online A/B Testing: Addressing Five Key Ques…	F4 (+F9)	ASO	R	AAUCCAUC
S-0033	J. A. Gagnon-Bartsch (2023). Precise unbiased estimation in randomized experiments using auxil…	F4	AOE	P	AAA–UUAC
S-0038	Olivier Jeunen (2026). Accelerating A/B-Tests with Counterfactual Estimation: Reducing V…	F7 (+F4)	AS	R	AUAA–UAC
S-0039	Olivier Jeunen (2026). Unifying On- and Off-Policy Variance Reduction Methods	F7 (+F4)	A	P	CCU—U–C
S-0040	Zhexiao Lin (2026). Variance reduction combining pre-experiment and in-experiment data	F4 (+F7)	AO	P	AAC–UUUC
S-0042	Sergei Pankratev (2026). Design-Aware Variance Reduction for Switchback Experiments: A Com…	F4	S	R	ACAC–CUC
S-0043	Xu Min (2026). Towards Reliable Social A/B Testing: Spillover-Contained Clusteri…	F2 (+F4+F5)	ASO	R	CCAAUUAC
S-0044	Tim Hesterberg (2024). Power Analysis for Experiments with Clustered Data, Ratio Metrics…	F4 (+F5)	AOE	R	ACU–UUUC
S-0046	David Arbour (2026). AI-Assisted Variance Reduction in Randomized Experiments	F4	ASOE	P	ACUCUCUC
S-0049	Neeti Pokharna (2026). Variance Reduction for Heavy-Tailed Monetization Metrics in Ranki…	F1 (+F4+F5)	SO	P	CCUUCCUC
S-0051	Yanping Pei (2026). Accounting for Classroom Effects in Student-Level Randomized Onli…	F4	ASOE	P	ACACCUUC
S-0056	Dae Woong Ham (2026). Design-Based Confidence Sequences: A General Approach to Risk Mit…	F9 (+F4)	ASO	P	AAACUAUC

Continued on next page

ID	Work	Family	Ev.	Pub	Appraisal
S-0060	Deddy Jobson (2024). Covariate Ordered Systematic Sampling as an Improvement to Random…	F1	ASO	P	CCAACUUC
S-0061	C. H. Bryan Liu (2023). An Evaluation Framework for Personalization Strategy Experiment D…	F1 (+F7)	AS	P	AC–A–CAC
S-0066	Murali Krishna Pasupuleti (2023). GenAI-Assisted ETL on Hadoop: Petascale Batch Causal A/B Pipelines	F4	S	P	CCUCCUUC
S-0067	Naoki Chihara (2026). Modeling Covariate Transition for Efficient Estimation of Longitu…	F8 (+F4)	ASO	U	AAAACAAC
S-0086	Qiong Zhang (2025). Rerandomization Algorithms for Optimal Designs of Network A/B Tests	F2	ASE	P	AA–ACUAC
S-0097	Adam C. Sales (2023). Using Auxiliary Data to Boost Precision in the Analysis of A/B Te…	F4	AOE	P	ACA–CUAC
S-0100	Ron Berman (2023). Latent Stratification for Incrementality Experiments	F5 (+F7)	ASO	P	ACAACCAC
S-0126	David Holtz (2020). Limiting Bias from Test-Control Interference in Online Marketplac…	F2 (+F7)	ASE	R	AC–CCUUC
S-0134	Alagappan Shanmugam (2026). Counterfactual Replay with Interference-Aware Stratification: A M…	F7 (+F2)	AS	R	CCUC–CUC
S-0146	Zhenghao Zeng (2026). Sequentially-Rerandomized Switchback Experiments	F3	ASE	R	AAAA–AUC
S-0188	Jinglong Zhao (2024). Pigeonhole Design: Balancing Sequential Experiments from an Onlin…	F1	ASE	P	AUCACUUC
S-0190	Sumin Shen (2023). Clustering-based Imputation for Dropout Buyers in Large-scale Onl…	F5 (+F4)	ASO	P	CCUACUUC
S-0431	Alex Deng (2015). Diluted Treatment Effect Estimation for Trigger Analysis in Onlin…	F7 (+F4)	AO	P	ACC–CUUC
S-0432	Alexey Drutsa (2015). Future User Engagement Prediction and Its Application to Improve…	F6	AOE	P	CCA–CCUC
S-0436	Keyu Nie (2020). Dealing With Ratio Metrics in A/B Testing at the Presence of Intr…	F5	ASE	P	CCUAUUUC
S-0440	Nian Si (2024). Tackling Interference Induced by Data Training Loops in A/B Tests…	F8 (+F7)	AS	R	ACAA–CUC
S-0466	Dominic Coey (2019). Improving Treatment Effect Estimators Through Experiment Splitting	F5 (+F9)	AO	P	ACAACUUC
S-0481	Qianyi Chen (2026). Journey to the Centre of Cluster: Harnessing Interior Nodes for A…	F2 (+F4+F7)	ASE	P	ACCACUAC
S-0524	Tomu Hirata (2025). Efficient and Scalable Estimation of Distributional Treatment Eff…	F5 (+F4)	ASO	R	ACAAAUAC
S-0564	Ziyad Benomar (2026). Augmented Hypothesis Testing with Persona-Based LLM Simulations	F4 (+F9)	ASE	R	UAUACCUC
S-0681	Alex Deng (2023). Variance Reduction Using In-Experiment Data: Efficient and Target…	F7 (+F6)	AO	P	CCC–CUUC
S-0719	Matt Taddy (2016). Scalable semiparametric inference for the means of heavy-tailed d…	F5	ASOE	P	ACAACUUC
S-0725	Alessandro Zito (2025). Pareto Optimal Proxy Metrics	F6	AO	P	AUA–CUUC

Continued on next page

ID	Work	Family	Ev.	Pub	Appraisal
S-0786	Ting Li (2023). Evaluating Dynamic Conditional Quantile Treatment Effects with Ap…	F8 (+F5)	ASO	P	CC–AUAUC
S-0795	Susan Athey (2023). Semiparametric Estimation of Treatment Effects in Randomized Expe…	F5	ASE	P	CCCAACAC
S-0802	Gabija Staponaitė (2025). Variance Reduction in Online Marketplace A/B Testing	F4 (+F5)	SE	P	CCUUCUUC

Table A2. The 46 contextual sources.

ID	Source	Year	Role
C-BC001	Winston Lin (2013). Agnostic notes on regression adjustments to experimental data:…	2013	Statistical foundation
C-BC013	Filip Radlinski (2013). Optimized Interleaving for Online Retrieval Evaluation	2013	Contextual, extracted
C-BC032	Jennifer Brennan (2025). Reducing Symbiosis Bias through Better A/B Tests of Recommenda…	2025	Contextual, extracted
C-BC050	David A. Freedman (2008). On regression adjustments to experimental data	2008	Statistical foundation
C-BC088	Not extracted for this contextual source (2020). Improving experimental power through control using predictions…	2020	Industry context
C-BC164	Patrick Bajari (2021). Multiple Randomization Designs	2021	Contextual, extracted
C-BC190	Brian Brost (2016). An Improved Multileaving Algorithm for Online Ranker Evaluation	2016	Contextual, extracted
C-BC191	Not extracted for this contextual source (2008). How does clickthrough data reflect retrieval quality?	2008	Contextual, boundary record
C-BC192	Not extracted for this contextual source (2014). Multileaved comparisons for fast online evaluation	2014	Contextual, boundary record
C-BC193	Harrie Oosterhuis (2017). Sensitive and Scalable Online Evaluation with Theoretical Guar…	2017	Contextual, extracted
C-BC195	Giuseppe Burtini (2015). A Survey of Online Experiment Design with the Stochastic Multi…	2015	Contextual, extracted
FND-FREEDMAN2008B	David A. Freedman (2008). On regression adjustments in experiments with several treatments	2008	Statistical foundation
S-0013	Tao Xiong (2021). Covariance Estimation and its Application in Large-Scale Onlin…	2021	Contextual, boundary record
S-0014	Pavel Serdyukov (2017). Machine Learning Powered A/B Testing	2017	Contextual, boundary record
S-0015	Ron Kohavi (2009). Controlled experiments on the web: survey and practical guide	2009	Predecessor review
S-0034	Matthew Gershoff (2025). K-Anonymous A/B Testing	2025	Contextual, boundary record
S-0035	Artur Markov (2024). Improved A/B testing acceleration methods for parametric hypot…	2024	Contextual, extracted
S-0037	Márton Trencséni (2024). The Unreasonable Effectiveness of Monte Carlo Simulations in A…	2024	Contextual, boundary record

Continued on next page

ID	Source	Year	Role
S-0064	Rahul Jain (2026). Foundations of A/B Experimentation Quality in Large-Scale Dist…	2026	Contextual, boundary record
S-0065	Peng Sun (2026). BoRP: Bootstrapped Regression Probing for Scalable and Human-A…	2026	Contextual, boundary record
S-0077	Wenxin Zhang (2025). Efficient Statistical Estimation for Sequential Adaptive Exper…	2025	Contextual, extracted
S-0437	Min Liu (2019). Large-Scale Online Experimentation with Quantile Metrics	2019	Contextual, extracted
S-0442	Alex Deng (2021). The equivalence of the Delta method and the cluster-robust var…	2021	Contextual, extracted
S-0470	Lorenzo Masoero (2024). Improved Prediction of Future User Activity in Online A/B Testing	2024	Contextual, extracted
S-0478	Venkat Kalyan Uppala (2022). From Data to Insight: Enhancing Content Creation with Advanced…	2022	Contextual, boundary record
S-0488	Akshay Balsubramani (2016). Sequential Nonparametric Testing with the Law of the Iterated…	2016	Contextual, extracted
S-0530	Rakibul Hasan (2023). A SYSTEMATIC REVIEW OF PREDICTIVE ANALYTICS IN MARKETING DECIS…	2023	Contextual, boundary record
S-0535	Mario Beraha (2026). Online activity prediction via generalized Indian buffet proce…	2026	Contextual, extracted
S-0561	Avinash Amudala (2026). System and Method for Reliability Scoring of Proxy Metrics in…	2026	Contextual, boundary record
S-0624	Federico Quin (2023). A/B Testing: A Systematic Literature Review	2023	Predecessor review
S-0626	Patrick Bajari (2023). Experimental Design in Marketplaces	2023	Predecessor review
S-0696	Wenxuan Guo (2025). ML-assisted Randomization Tests for Detecting Treatment Effect…	2025	Contextual, boundary record
S-0729	Prasanjit Dubey (2026). BOOST: Power-Optimal Strong-FWER Testing for Block-Structured…	2026	Contextual, boundary record
S-0740	Alexey Kurennoy (2025). YEAST: Yet Another Sequential Test	2025	Contextual, boundary record
S-0781	Steven R. Howard (2022). Sequential estimation of quantiles with applications to A/B te…	2022	Contextual, extracted
S-0782	Weinan Wang (2020). CONQ: CONtinuous Quantile Treatment Effects for Large-Scale On…	2020	Contextual, extracted
S-0783	Yuxiang Xie (2020). How to Measure Your App: A Couple of Pitfalls and Remedies in…	2020	Contextual, extracted

Continued on next page

ID	Source	Year	Role
S-0784	Mårten Schultzberg (2022). Resampling-free bootstrap inference for quantiles	2022	Contextual, extracted
S-0785	Michael Lindon (2025). Anytime-Valid Inference in Linear Models and Regression-Adjust…	2025	Contextual, extracted
S-0787	Leon Yao (2024). Privacy-preserving Quantile Treatment Effect Estimation for Ra…	2024	Contextual, extracted
S-0794	Leheng Cai (2026). Asymptotic Anytime-Valid Quantile Inference under Local Differ…	2026	Contextual, boundary record
S-0797	Evan Miller (2023). Likelihood-ratio inference on differences in quantiles	2023	Contextual, boundary record
S-0799	Zhaoqi Li (2026). Estimating the True Effect Size Distribution with SIMEX	2026	Contextual, boundary record
S-0800	Nicholas Larsen (2024). Statistical Challenges in Online Controlled Experiments: A Rev…	2024	Predecessor review
S-0801	Iavor Bojinov (2022). Online Experimentation: Benefits, Operational and Methodologic…	2022	Predecessor review
S-0803	Florian Auer (2021). Controlled Experimentation in Continuous Experimentation: Know…	2021	Predecessor review

Appendix B. Search queries and retrieval totals

The five final OpenAlex queries used the works endpoint with the combined title-and-abstract search field; the three final arXiv queries used the export API over all fields. Raw responses, exact request parameters, timestamps, and pagination states were saved before any filtering. Totals (N) are query occurrences before cross-query deduplication; OpenAlex returned 975 occurrences representing 896 distinct source records and arXiv 85 occurrences representing 66, giving 1,060 occurrences and 962 records.

Table B1. Final query strings and retrieval totals.

ID	Source	Query string (verbatim)	N
OA-TA-01	OpenAlex	(“online controlled experiment” OR “online controlled experiments” OR “A/B testing” OR “A/B tests” OR “online experiments”) AND (“variance reduction” OR “covariate adjustment” OR “control variates” OR CUPED OR CUPAC)	100
OA-TA-02	OpenAlex	(“online controlled experiment” OR “online controlled experiments” OR “A/B testing” OR “A/B tests” OR “online experiments”) AND (stratification OR “stratified randomization” OR rerandomization OR “re-randomization” OR “matched pairs” OR blocking OR “balanced allocation”)	411
OA-TA-03	OpenAlex	(“online controlled experiment” OR “online controlled experiments” OR “A/B testing” OR “A/B tests” OR “online experiments”) AND (“regression adjustment” OR “machine learning” OR prediction OR “ratio metrics” OR “quantile metrics” OR “delta method”) AND (variance OR precision OR sensitivity OR power)	268

Continued on next page

ID	Source	Query string (verbatim)	N
OA-TA-04	OpenAlex	(“online controlled experiment” OR “online controlled experiments” OR “A/B testing” OR “A/B tests” OR “online experiments”) AND (review OR survey OR overview) AND (“variance reduction” OR “statistical methods” OR “experimental design”)	68
OA-TA-05	OpenAlex	(“online experimentation” OR “AB testing” OR “A/B experiment” OR “A/B experiments”) AND (“variance reduction” OR “covariate adjustment” OR “control variates” OR CUPED OR CUPAC OR stratification OR “stratified randomization” OR rerandomization OR “re-randomization” OR “matched pairs” OR blocking OR “balanced allocation” OR “regression adjustment” OR “ratio metrics” OR “quantile metrics” OR “delta method” OR ((“machine learning” OR prediction) AND (variance OR precision OR sensitivity OR power)) OR ((review OR survey OR overview) AND (“statistical methods” OR “experimental design”)))	128
AX-01	arXiv	(all:“online experiments” OR all:“online controlled experiments” OR all:“A/B testing”) AND (all:“variance reduction” OR all:“covariate adjustment” OR all:“CUPED” OR all:“CUPAC”)	37
AX-02	arXiv	(all:“online experiments” OR all:“online controlled experiments” OR all:“A/B testing”) AND (all:“stratification” OR all:“rerandomization” OR all:“ratio metrics” OR all:“quantile” OR all:“regression adjustment”)	36
AX-03	arXiv	(all:“online experimentation” OR all:“AB testing” OR all:“A/B experiment” OR all:“A/B experiments”) AND (all:“variance reduction” OR all:“covariate adjustment” OR all:“control variates” OR all:“CUPED” OR all:“CUPAC” OR all:“stratification” OR all:“rerandomization” OR all:“ratio metrics” OR all:“quantile” OR all:“regression adjustment”)	12