BRUSHWORKS · METHODOLOGY NOTES · VERSION 1.0 · SEPTEMBER 2026

Methodology notes.

The standard of proof, written down

How every Brushworks program is designed to be measured, how its results are computed, and what may and may not be reported: the sealed holdout, register arms, the intention-to-treat estimand, power, the publish gate, the daily loop, decay and backfire, reporting standards and limits – with a worked example and templates.

Brushworks

brwrks.com/reports
research@brwrks.com
Cite as: Brushworks (2026), *Methodology notes*, Version 1.0 · September 2026.

Contents

Section 1 Purpose, scope and standing

Section 2 Definitions

Section 3 The sealed holdout

Section 4 Register-level arms

Section 5 Outcomes

Section 6 Estimand and estimation

Section 7 Power and minimum detectable effects

Section 8 The publish gate

Section 9 The daily loop

Section 10 Decay and backfire

Section 11 Reporting standards

Section 12 Limits and known failure modes

Section 13 Independent review

Appendix A A worked example

Appendix B Balance table template

Appendix C Model card template

Appendix D References

Section 1

Purpose, scope and standing

These notes set out how every Brushworks program is designed to be measured, how its results are computed, and what may and may not be reported. They are the standard the Brushworks Review cites, the standard every field report and proof report is written to, and the standard an independent reviewer assesses each year. A program that cannot be run to these notes is not accepted.

The notes apply to candidate, independent-expenditure, ballot-measure, bond and institutional programs alike. Where an institutional mandate has no electoral outcome, the recorded acts in Section 5 are replaced by the mandate's own registered outcomes (enrollment, uptake, attendance, compliance behavior) and the rest of the notes apply unchanged. Where the notes are silent, the default is the more conservative choice, stated in the report.

THE ONE SENTENCE

Ten percent of every cohort is drawn at random before any money is spent, sealed, never contacted, and reported to the client directly; every claim of effect is the difference between the people we treated and the people we left alone, with its uncertainty stated.

1.1 How to use this document

Section 2 defines the terms. Sections 3 to 6 describe the design: the holdout, the register arms, the outcomes, and the estimand. Section 7 gives the power tables the terrain brief uses to size a program. Sections 8 to 10 govern the models and the daily loop. Section 11 fixes what a report contains. Section 12 lists the limits the firm will state rather than hide. The appendices carry a worked example, templates and a glossary. A campaign manager needs Sections 3, 7 and 11; a data buyer needs 8 and 12; counsel needs 3.4 and 11.4.

1.2 Versioning and review

This is version 1.0, September 2026. The notes are reviewed annually by an independent reviewer with field-experimental expertise and after any program in which the design failed; revisions are posted at brwrks.com/reports with a new version number and a dated change log. A program is measured to the version in force on the day its holdout was sealed.

Section 2

Definitions

Program
A bounded engagement with one client, one purpose, one file and one measurement design, from terrain brief to proof report.

Cohort
A group of individuals on the segment map who share what would move them, drawn from behavior; named as a composite (the Steward, the Night Nurse) and sized before design.

Register
One of three disclosed channels of persuasion: precision paid (the sponsor's own voice, placed at the person), organic advocacy (a real, disclosed second voice), and ambient and earned (the argument from the world).

Holdout (control)
The random ten percent of each cohort assigned on day five to receive nothing from the program in any register.

Treated
Everyone in a cohort not in the holdout, whether or not the program actually reached them.

Lift
The difference in an outcome between treated and holdout, in percentage points, with its interval.

Estimand
The quantity a report estimates: the intention-to-treat effect of assignment to the program on the outcome, in the population of the cohort.

Interval
A ninety-five percent confidence interval, two-sided, computed as Section 6 describes.

Baseline
A model a candidate model must beat: the voter file used well, and the platforms' look-alike targeting (Section 8).

Publish gate
The pre-registered test a model passes before it is used in a program.

Kill threshold
The daily in-field discrimination level below which a shipped model is retired.

Decay
The decline of a measured effect over time after contact.

Backfire
A negative lift: contact that moved a cohort away from the outcome.

Contamination
Contact of holdout members by the program's own registers (a design failure) or by other actors (a limit to be reported).

Recorded act
An outcome observed in an administrative or transactional record with a timestamp: a validated vote, a returned ballot, a contribution, a sign-up.

Panel measure
An outcome measured by the same instrument administered to treated and holdout at the same times, used when no act exists.

Section 3

The sealed holdout

3.1 Sampling frame

The frame is the segment map: the client's file matched to the graph, scored, separated into cohorts, with the hold list (people who receive nothing regardless) removed before sampling. Individuals are the unit; where a source attributes behavior to the household rather than the person, the household is the unit and every member is assigned together. The frame is frozen on day five and hashed; the hash is delivered to the client with the assignment list.

3.2 Assignment

Within each cohort, a simple random sample of ten percent is assigned to the holdout using a seeded generator whose seed is recorded in the program file. Stratification by cohort is the only stratification; within a cohort, no covariate is used to assign, so that balance is a property of chance to be checked rather than of design to be argued. A cohort smaller than two thousand is pooled with its nearest neighbor on the map for the purpose of measurement, and the report says which cohorts were pooled and why.

3.3 Sizing

Ten percent is the default because a statewide program with a hundred thousand treated voters yields cohort holdouts of several hundred to a few thousand, which Section 7 shows can see the effects that matter at the cohort level, while leaving the program ninety percent of its reach. A client may elect a larger holdout for precision; the firm will not run a smaller one, because a holdout that cannot see a three-point effect at the program level cannot fulfil Section 11.

3.4 Sealing and delivery

The assignment list is hashed with a keyed function and delivered to the client directly, outside the firm's delivery path, on day five. The firm's seats receive the holdout only as a suppression list. No member of the delivery team can see which individuals are in the holdout; the measurement team can. The holdout is never contacted by the program in any register for the life of the program, and is not re-drawn if the segment map is revised; new cohorts created mid-program receive a new holdout drawn the same way.

3.5 Balance and integrity checks

On day six the measurement team produces a balance table (Appendix B): for each cohort, the means of pre-treatment covariates — participation history, age band, prior-cycle contact, the cohort's own score distribution — in treated and holdout, with standardized differences. A standardized difference above 0.1 on any covariate triggers a re-draw before any spend. During the program, contamination is monitored two ways: every placement log is matched against the holdout hash and any match is an incident; and, where the client can supply its other vendors' contact records, holdout contact by those vendors is recorded and reported as a limit under Section 12.

What invalidates a holdout. Drawing it after any spend; contacting it in any register; sizing it below the threshold in Section 7 without stating the resulting minimum detectable effect; re-drawing it after outcomes have begun to accrue; or replacing it with a matched comparison. A program in which any of these occurs is reported as unmeasured, and the client's fee for measurement is not charged.

Section 4

Register-level arms

Where the treated population of a cohort exceeds twenty thousand, it is further divided at random into arms that receive precision paid only; paid and organic; and all three registers in sequence. The differences between arms estimate each register's marginal contribution; the difference between the all-register arm and the sum of the single-register contributions estimates the interaction, which Review paper WP-07 reports as the value of sequencing. Where the population does not allow arms, all registers are measured together against the single holdout and the report states that register contributions are not separately identified.

Arms are assigned on day five with the holdout, from the same seed, and sealed the same way. A register that is added mid-program is measured by a new arm only if the remaining program is long enough for Section 7's power; otherwise its effect is reported as pooled.

Section 5

Outcomes

5.1 The hierarchy

RANK	OUTCOME	SOURCE	USED FOR
1	Validated turnout	The state's post-election voter history	Mobilization programs; available only after certification
2	Ballot returned	State or county absentee and early-vote files, daily	Mobilization, in flight, in vote-by-mail states
3	Contribution, sign-up, attendance, volunteer shift	The client's own transactional records	Engagement and fundraising programs
4	Registered institutional outcome	The mandate's administrative records	Institutional programs
5	Favorability, intent, issue position	Panel measurement (5.2)	Persuasion programs, where no act exists

A program's primary outcome is chosen from the highest available rank and pre-registered in the terrain brief. Secondary outcomes are named there too. Self-reported turnout is never an outcome, because it is inflated in ways that differ between treated and untreated.

5.2 Panel measurement

Where the primary outcome is an attitude, a panel is drawn from treated and holdout in the same proportions, measured with the same instrument, by the same mode, at the same times, with the same incentives, and without any indication of assignment. Panel attrition is reported by arm; differential attrition above five points triggers a sensitivity analysis under Section 6.5. The instrument is fixed before the program and is not altered in flight.

Section 6

Estimand and estimation

6.1 Intention to treat

The estimand is the effect of assignment to the program, not of contact: a treated voter never reached counts as treated. This is the only estimand a campaign can act on, because the campaign controls assignment and not attention; it is also the conservative one, since non-contact dilutes the effect toward zero.

6.2 Estimation

Lift is the difference in means (or proportions) between treated and holdout within a cohort, or within an arm. Regression adjustment for the pre-treatment covariates in the balance table is used to improve precision and is reported alongside the unadjusted difference; the covariates are those pre-registered in the terrain brief and no others. Where assignment was at the household, standard errors are clustered at the household.

6.3 Intervals

Intervals are ninety-five percent and two-sided. For proportions with cohort holdouts above four hundred, the normal approximation is used; below that, an exact interval. Program-level lift is the cohort-weighted average of cohort lifts, weighted by cohort size, with its interval computed from the cohort variances.

6.4 Multiple comparisons

The primary outcome in the primary population is reported without adjustment. Secondary outcomes, register arms and cohort-by-frame cells are reported with a false-discovery-rate adjustment across the family, and the report states the number of comparisons. A cohort's lift is described as "not distinguishable from zero" when its adjusted interval includes zero, never as "no effect".

6.5 Missing data and attrition

Recorded acts have no missingness by construction. Panel attrition is handled by inverse-probability weighting on pre-treatment covariates, and the report presents bounds under the assumption that attriters in the two arms differ by the observed differential.

Section 7

Power and minimum detectable effects

The terrain brief states, before a program starts, which effects its design can see. The tables assume eighty percent power, a two-sided five percent test, and a treated group at least as large as the holdout (which the ten-percent rule guarantees). Minimum detectable lift is in percentage points.

HOLDOUT SIZE	BASE RATE 10%	BASE RATE 30%	BASE RATE 50%	DESIGN IMPLICATION
500	3.7	5.6	6.1	Pool cohorts; program-level reading only
1,000	2.6	4.0	4.3	Pool small cohorts; registers together
2,000	1.9	2.8	3.1	Cohort-level lift; registers pooled
5,000	1.2	1.8	2.0	Cohort and register lift separately
10,000	0.8	1.3	1.4	Frame-by-cohort tests in the second month
20,000	0.6	0.9	1.0	Decay and backfire watched daily with confidence

Table 7.1 Minimum detectable lift by holdout size and base rate. Values assume the treated arm is at least the holdout's size; larger treated arms improve precision modestly.

Rules of thumb that follow: a statewide program of a hundred thousand treated voters (ten thousand held out) can see a one-and-a-half-point program effect and three-to-four-point cohort effects; a congressional program of twenty thousand can see a three-point program effect and only pooled cohort effects; a municipal program below five thousand is measured at the program level only, and the terrain brief says so. A program that is underpowered for its client's question is redesigned (larger universe, pooled cohorts, a longer field period) or the question is restated; it is not run and then reported as if it could see what it could not.

Section 8

The publish gate

8.1 Baselines

The file used well. A logistic model of the outcome on participation history (last four general and primary elections), party registration, age band, precinct-level partisanship and prior-cycle contact — what a competent consultant builds from the voter file alone. **The look-alike.** The delivery platforms' own audience expansion seeded with the campaign's known supporters at the platforms' default settings, scored on the client's holdout by matched exposure. Both baselines are built for every program, and their discrimination is reported in the proof report whether or not a candidate model beat them.

8.2 The test

A candidate model is trained on recorded outcomes from prior programs and validated on a time-split: trained on cycles before a cutoff, tested on the client's own holdout for the current program's outcome. Discrimination is reported as area under the curve and as lift in the top decile. The gate is a pre-registered margin over the stronger baseline — by default 0.03 in AUC and ten percent relative in top-decile lift — with the margin stated in the model card. A model that fails the gate is retired and recorded in the kill log with the baseline that beat it.

8.3 Proxy reconstruction

Every candidate model is tested for reconstructing an excluded attribute — ethnicity, religion, ideology — by training a classifier on the model's outputs against a labeled reference set. A classifier achieving an AUC above 0.60 on any excluded attribute retires the model regardless of performance, and the kill log records it under proxy reconstruction.

8.4 Model cards

Every shipped model carries a card (Appendix C) stating its family, training basis, validation split, baselines and margins, the outcome it predicts, the cohorts it may score, its known limits, the proxy test result, its field kill threshold, and its owner. The card is delivered to the client with the segment map.

8.5 The field kill threshold

In field, a shipped model's discrimination is recomputed daily on the accruing outcomes of the treated arm against its holdout. A model whose seven-day rolling AUC falls below its card's threshold — by default the stronger baseline's AUC — for three consecutive days is retired, the cohort is rescored by the next model on the card or by the file, the client is told in that week's report, and the kill log records the date and the drift measure.

Section 9

The daily loop

Every day a program is in field, the measurement team reads four things by cohort and register: the treated-minus-holdout difference on the leading indicator (returns, panel favorability, response), its sign, its trend over seven days, and each model's discrimination. Three decisions follow rules written before the program starts:

SIGNAL	RULE	ACTION	RECORDED
Reflight	A cohort's register effect exceeds its seven-day average and its interval excludes zero	Weight in that register raised for that cohort within the approved plan	Reallocation log; weekly report
Suppress	A cohort's register effect is negative for three consecutive readings	That register suppressed for that cohort; cohort re-examined; if negative in every register, cohort moves to the hold list	Suppression log with the interval; weekly report
Retrain	A model breaches its kill threshold (8.5)	Model retired; cohort rescored; next model or the file used	Kill log; weekly report

Reallocations above ten percent of any budget line require the client's written approval under the governance policy; the loop proposes, the client disposes.

Section 10

Decay and backfire

10.1 Decay

Persuasion effects decay; the literature places the half-life of televised advertising effects at days to a week or two, and the firm's own programs show cohort-dependent half-lives from one to five weeks. Decay is measured by the treated-minus-holdout difference over time after a flight and reported as a half-life per cohort and register where the design has the power; the plan uses decay to schedule reflights so that effects are present at the ballot, not at the peak of spend.

10.2 Backfire

A negative lift with an interval excluding zero is backfire. It is reported in the same table as positive lift, with the same prominence, by cohort and register, together with the suppression decision and its date. The proof report never omits a cohort that moved the wrong way. Review paper WP-26 sets out what produces it; these notes require only that the design can see it and the report shows it.

Section 11

Reporting standards

11.1 The weekly field report

ITEM	CONTENT
Lift table	By cohort and register: treated-minus-holdout on the leading indicator, adjusted and unadjusted, with intervals and the number of comparisons
Loop decisions	Reflights, suppressions and retrains this week, each with the rule that triggered it
Models	Each shipped model's seven-day discrimination against its threshold
Integrity	Holdout contamination checks; balance re-check if the map changed
Media trail	Spend by register and platform from the accounts of record; placements visible in ad libraries
Plan	Next week's weights, with reallocations needing approval flagged

11.2 The proof report

SECTION	CONTENT
Design	Holdout size by cohort, seeds, balance table, arms, pre-registered outcomes and covariates, the version of these notes
Results	Primary outcome: program lift and cohort lifts with intervals; secondary outcomes with adjustment; register contributions and interaction where identified
Decay and backfire	Half-lives where measured; every negative lift with its suppression date
Models	Model cards; the program's kill log; baseline discrimination
Media trail	Invoice to placement to impression, by register, reconciled to the accounts of record
Limits	Every item in Section 12 that applied
Basis	Which figures are the program's, which are pooled, which are illustrative; rounding

11.3 Rounding and labeling

Lift is reported to one decimal place; intervals to one decimal place; discrimination to two. Every figure in a public document is labeled official (with its source), the firm's record, pooled and rounded across programs, or illustrative of the method. Program figures are pooled and rounded in public and the client is never identified.

11.4 What is never reported

- Raw outcomes among the treated presented as an effect of the program.
- The race's result presented as the program's result, or vote share that coincided with the program.

- Lift without its interval, or a cohort's negative lift omitted.
- Effects the design could not see under Section 7, presented as effects.
- Self-reported turnout as an outcome.

Section 12

Limits and known failure modes

LIMIT	WHAT IT DOES	HOW IT IS HANDLED AND REPORTED
Contamination by others	The campaign's other vendors, the opponent and the news reach holdout members	Reported where the client can supply contact records; interpreted as the program's effect over and above the environment both arms share
Spillover	Treated voters talk to holdout members in the same household or workplace	Household assignment removes within-household spillover; residual spillover biases toward zero and is stated
Non-contact	Many treated voters are never reached	The intention-to-treat estimand absorbs it; the report gives reach by register so the client can judge the per-contact effect
Platform measurement gaps	Some platforms do not return exposure at the individual level	Assignment, not exposure, defines arms, so estimation is unaffected; the media trail states which platforms report exposure
Small races	Universes too small for cohort-level power	Program-level measurement only, stated in the terrain brief before engagement
Late outcomes	Validated turnout arrives weeks after the election	A preliminary proof report on leading indicators; a final report when the state's history is published
Instrument effects	Panel measurement changes attitudes	Identical instruments in both arms cancel the effect on the difference; the level is not interpreted

Section 13

Independent review

An independent reviewer with field-experimental expertise examines a sample of proof reports and the program files behind them each year, checks the holdout seeds, balance tables and estimation code against these notes, and publishes a summary with the annual report. The reviewer's letter is available to clients on request. Deviations found are reported with the affected programs identified to the clients concerned and described in aggregate publicly.

Appendix A

A worked example

A statewide persuasion program with 118,000 voters on the segment map after the hold list is removed, six cohorts, a nine-week field period, and a panel primary outcome of favorability. Illustrative of the method; figures rounded.

COHORT	SIZE	HOLDOUT	MDE (PTS)	LIFT AT WEEK 8	INTERVAL	READING
The Steward	31,400	3,140	2.5	+3.9	2.6 to 5.2	Moved; reflighted week 6
The Night Nurse	22,800	2,280	2.9	+4.6	3.2 to 6.0	Moved; largest effect
The Renter	19,600	1,960	3.1	+3.1	1.7 to 4.5	Moved
The Splitter	17,900	1,790	3.3	+1.8	0.4 to 3.2	Moved, small; late-deciding
The Pharmacist	14,100	1,410	3.7	+0.9	−0.6 to 2.4	Not distinguishable from zero
The Contractor	12,200	1,220	4.0	+2.7	1.1 to 4.3	Moved
Program	**118,000**	**11,800**	**1.3**	**+3.1**	**2.4 to 3.8**	**Cohort-weighted; primary outcome**

Table A.1 Design and result. The Certain (9,400) were on the hold list and outside the frame. The Pharmacist's interval includes zero and is reported as such, not as "no effect".

Register arms were run in the three cohorts above twenty thousand. In the Steward, precision paid alone moved 2.4 points, paid and organic 3.3, and all three in sequence 4.5; the interaction of 0.8 points is the value of sequencing, with an interval of −0.3 to 1.9 — suggestive, not established, and reported as such. Two models were retired in field: a dispositions model whose AUC fell from 0.71 to 0.63 in week four (rescored by the file at 0.66), and an interests model retired at the gate before the program for proxy reconstruction of religion at AUC 0.64. The media trail reconciled $2.31 million of placements to the accounts of record with no unexplained variance.

Appendix B

Balance table template

COVARIATE	TREATED MEAN	HOLDOUT MEAN	STD. DIFFERENCE	THRESHOLD
Voted in last general	0.71	0.70	0.02	≤ 0.10
Voted in last primary	0.38	0.39	−0.02	≤ 0.10
Age band 45–64	0.34	0.34	0.00	≤ 0.10
Contacted last cycle	0.52	0.53	−0.02	≤ 0.10
Cohort score, mean	0.61	0.61	0.01	≤ 0.10

Table B.1 Produced per cohort on day six; any standardized difference above 0.10 triggers a re-draw before spend.

Appendix C

Model card template

FIELD	CONTENT
Identifier and family	e.g., D-07, dispositions
Outcome predicted	The recorded act or panel measure, and the population
Training basis	Programs and cycles used; time-split cutoff
Validation	Client holdout; AUC and top-decile lift; baselines' AUC; margin over the stronger baseline
Proxy test	Classifier AUC on each excluded attribute; pass threshold 0.60
Permitted use	Cohorts and registers the model may score; jurisdictions
Known limits	Populations or periods where discrimination is weaker
Field kill threshold	Seven-day AUC below which the model is retired; three-day rule
Owner and date	Head of research; card date; retirement date if any

Appendix D

References

- Gerber, A. S., and Green, D. P. (2012). *Field Experiments: Design, Analysis, and Interpretation*. W. W. Norton.
- Kalla, J. L., and Broockman, D. E. (2018). The minimal persuasive effects of campaign contact in general elections. *American Political Science Review*, 112(1), 148–166.
- Coppock, A., Hill, S. J., and Vavreck, L. (2020). The small effects of political advertising are small regardless of context, message, sender, or receiver. *Science Advances*, 6(36).
- Gerber, A. S., Gimpel, J. G., Green, D. P., and Shaw, D. R. (2011). How large and long-lasting are the persuasive effects of televised campaign ads? *American Political Science Review*, 105(1), 135–150.
- Gordon, B. R., Zettelmeyer, F., Bhargava, N., and Chapsky, D. (2019). A comparison of approaches to advertising measurement. *Marketing Science*, 38(2), 193–225.
- Lewis, R. A., and Rao, J. M. (2015). The unfavorable economics of measuring the returns to advertising. *Quarterly Journal of Economics*, 130(4), 1941–1973.
- Mitchell, M., et al. (2019). Model cards for model reporting. *Proceedings of FAT* 2019*.
- Brushworks Review WP-06, WP-20, WP-25, WP-26.

Methodology notes, version 1.0, September 2026. Independent review: annual. Change log: initial publication.