Method Disclosure in Published Marketing Results

A Cross-Sectional Analysis of 6,998 Published Case Studies from 369 Media-Placing Firms

Trevor Birba

Dark Wave Marketing Science

28 August 2026

darkwavemarketing.science/notes/method-disclosure-in-case-studies

Contents

Abstract

658 marketing services websites were crawled and parsed, yielding 18,459 numeric claims of a marketing result. The analysis population is the 369 of those firms that place media on a client's behalf, whose sites carry 6,998 published case studies and 15,662 numeric claims; 3,386 of those case studies publish at least one numeric claim and carry 13,853 of them. Every case study was classified for the presence of method language on the page, and every claim for four properties a reader requires in order to interpret a reported percentage.

Most firms never publish evidence of a comparison condition. 256 of 369 firms (69.4%) show no comparison of any kind anywhere on their site. 62 firms (16.8%) show testing discipline without a named design. Only 51 firms, 13.8% of the population, ever describe running a named experimental design. At the level of the individual case study, 2,941 of 3,386 published case study pages (86.9%) show no comparison condition at all.

Capability does not produce disclosure. The 113 firms that describe any applied method still publish 81.2% of their case study pages with no method stated. Narrowing to the

51 firms demonstrably running named designs, 77.8% of their case study pages, and 76.6% of their 4,820 published numeric claims, appear with no mention of a method in any form — not the design, not a comparison group, not a statistical test, not a measurement partner, not a period over which the comparison was drawn. These firms are their own control, so the publication threshold governing which results a firm elects to show operates identically on both sides of the comparison and cancels.

The rate is statistically indistinguishable between advertising agencies and the platforms and vendors operating alongside them, and, once the publishing firm is held constant, between client verticals. Disclosure is a property of the publishing firm, not of the client's industry or the firm's category. Between-firm variation accounts for 51.9% of page-level variance in disclosure; between-vertical variation accounts for 4.8%, and none of it survives within-firm comparison.

Nothing visible on the page substitutes for the missing method. Claim magnitude carries no information about disclosure: disclosed and undisclosed results have median magnitudes of 51% and 50%, and the within-firm ordering is indistinguishable from chance. A preregistered hypothesis that weaker methodological language would accompany larger claimed effects is rejected, which removes the one heuristic a reader might otherwise apply.

The four interpretive properties are near-universally absent, and uniformly so. After correction for classifier error, a numeric baseline is absent from 99.2% of claims, the source of the figure from 94.7%, the period from 84.9%, and enumerated concurrent activity from 82.4%. The spread across those four figures is the difference between the properties themselves, not an uncertainty band; each carries its own confidence interval. No subgroup is better served: the properties are absent from between 91% and 99% of claims in every vertical and at every tier of methodological capability.

Absent all four, a published percentage supports no inference. The reader cannot determine what the figure is a proportion of, over what interval it was accumulated, what else was running, or who produced it, and therefore cannot validate or discard the stated impact of the work. Four of the five conditions this paper sets out require only a sentence of additional text and no additional measurement capability. On that basis the gap is one of convention, not of capability.

The paper closes by stating the criteria the authors apply before standing behind a reported number, together with the inter-rater reliability obtained when independent readers were asked to apply them.

Competing interests. Dark Wave Marketing Science sells measurement services. A finding that measurement is poorly disclosed is self-serving on its face. The disclosure is placed first because it is the first objection a reader should raise. It is addressed by publishing the method, the instrument failures, and the results that ran against the authors' expectations.

1. Scope

This is not a study of the accuracy of published marketing claims, and cannot be. The underlying data behind a case study is not published, and in most cases no counterfactual appears to have been constructed: of the 3,386 published case study pages carrying at least one numeric claim, 2,941 (86.9%) show no comparison condition of any kind; 183 pages (5.4%) describe a named design and 262 (7.7%) describe testing without one. Where no comparison was constructed there is nothing to verify a claim against, and no external standard exists. Every result reported here concerns what a published page permits a reader to conclude, and none concerns whether a reported number is true.

The distinction is load-bearing, not defensive. A published number carrying no method is not a weakly supported causal claim; it is not a causal claim at all. It is an observation about a metric with no attribution attached. "Bookings rose 47%" and "we drove a 47% increase in bookings" are different propositions requiring different evidence, and only the second requires a method. The observation that motivates this study is that the industry publishes the first construction and is read as having asserted the second.

Every claim in the corpus was selected for publication by the firm that produced it, placed on a marketing page, under no client confidentiality constraint and with every incentive to demonstrate rigor. This makes every rate reported here a generous estimate of disclosure practice. A firm choosing which of its engagements to display on a marketing page will not lead with the ones it can say least about; the corpus is the industry's own selection of its most presentable work. The direction of the resulting bias is unambiguous and runs in one direction only: whatever disclosure rate is observed here is at least as high as, and probably higher than, the rate across all work performed. When 76.6% of the claims published by demonstrably capable firms carry no method, that is a floor on non-disclosure, not an estimate of it.

2. Method

2.1 Sampling frame

658 websites were crawled between 25 and 28 August 2026, identified through two independent paths. The first is a search grid constructed around result-claim language instead of firm names, on the reasoning that a page already publishing the phrase "return on ad spend" is itself the sampling unit of interest. The second is enumeration from the client side within a defined region and set of verticals, identifying the firm holding each account. This path reaches firms with no directory presence, the small end of the size distribution.

Inclusion criteria. A site enters the corpus if it returns usable text, evidences paid media execution on a client's behalf, and is not excluded by a named non-participant rule. Me-

dia execution is scored from explicit buying language ("we buy," "media plan," "campaign management," named platforms bought against) or, at a lower confidence, from paid channel execution without an explicit buying line. Both grades are retained; a site that explicitly disclaims media buying is excluded.

Exclusion criteria, applied by named rule and inspectable. Firms participating in no client advertising outcome are removed: business and trade press, industry directories, associations, research vendors, government bodies, destination marketing organizations acting as advertisers instead of agents, software vendors selling a product and not a service, and subsections of a larger organization crawled separately from their parent. Rules that could plausibly capture a real firm yield to an agency term in the domain, so that a firm named `impressivedigital` is not removed for containing the string `press`. Every exclusion is listed in Appendix B and every rule was tested against the full media-placing set for false removals before application.

Crawl scope per site. The home page, services pages, and all reachable case study, client work and results pages, discovered by sitemap where one is published and by link traversal otherwise, to a cap. Every judgment about a firm is made over all of its retrieved pages, never a subset; three separate defects in earlier versions of this instrument arose from judging a site on its home page alone, and the rule now sits in the code at each call site.

The frame is not a census. US marketing agencies number in the tens of thousands and no crawl of this size is representative of that population. Every rate reported below is a rate within the corpus and is stated as such.

	Count	Share
Websites crawled	658	
Returned usable text	552	83.9%
Claims retained	18,459	
Case studies published by the 369 participating firms	6,998	
…of which publish at least one numeric claim	3,386	48.4%
Claims held by participating firms	15,662	

Two further properties were recorded. They are independent of one another, not sequential filters: a firm may place media without publishing a numeric claim, and may publish claims without placing media, so the counts do not nest.

	Count	Share of 658
Places media	390	59.3%
Publishes at least one numeric claim	505	76.7%
Both, the analysis denominator	348	52.9%

A firm that does not place media has no media result to attribute, and is therefore outside the scope of the question, not excluded for analytic convenience.

2.2 Population definition

The unit of analysis is the firm that supports client advertising outcomes, not the advertising agency. Of the 658 sites, 406 classify as agencies and 252 as platforms and vendors; within the media-placing subset the split is 276 to 114. That line is not treated as a population boundary. An agency and a demand-side platform both place media on a client's behalf, both publish numeric results derived from it, and both face the same decision about whether to state how the number was produced. The relevant population is defined by that shared position, not by firm type.

Two consequences follow from adopting the wider definition.

First, a third category is excluded that does not belong under either definition. An audit of the media-placing subset established that the substantive contaminant was never platforms but firms participating in no client advertising outcome at all: business and trade press, industry directories, associations and research vendors, which enter a domain crawl because they publish on the same subject matter. Twenty-one of the 390 media-placing hosts are removed on this basis, leaving 369 participating firms, of which 271 classify as agencies and 98 as platforms and vendors. Every removal is listed in Appendix B.

Second, the classifier separating agencies from platforms ceases to be load-bearing. Under the narrower definition, a pooled figure had to be defended by a split, and the split depended on an unvalidated label. Under the wider definition the split becomes a robustness check. Section 3.2 reports that the two groups are statistically indistinguishable on every measure presented here.

2.3 Claim eligibility

A number appearing on a marketing services website is usually not a marketing result. Prior to any classification, an eligibility gate refused 15,534 numbers, each with a recorded reason:

Reason refused	Count	Share of refusals
Street addresses, ZIP codes, phone numbers	5,738	36.9%

Reason refused	Count	Share of refusals
Index-card republication of a detail page	2,965	19.1%
Fragment too short to interpret	1,779	11.5%
Bare five-digit integer with no unit	1,060	6.8%
Zero-value animated counter	639	4.1%
Navigation and calls to action	619	4.0%
Pricing tables and project-size bands	576	3.7%
Truncated quote	487	3.1%
Vanity totals and deliverable counts	446	2.9%
Cookie banners, copyright, boilerplate	403	2.6%
CSS and markup fragments	312	2.0%
Third-party industry statistics	291	1.9%
Client problem statements	139	0.9%
Audience targeting specifications	52	0.3%
Market context about the client's territory	28	0.2%

This table is published because the eligibility gate is the component of the instrument most likely to be wrong and the component a reader cannot otherwise inspect. One refused value was `404040`, extracted from the CSS fragment `text-[#404040]`.

2.4 Detection of method language

Method language is scored in two tiers. An A/B test on a landing page constitutes genuine testing discipline and is not a geographic holdout; collapsing the two produced misclassification in both directions.

Strong denotes a named design incorporating a comparison group: control group, holdout, matched market, synthetic control, geographic test, exposed and unexposed cells,

treatment groups, brand or conversion lift study, incrementality test, or an identified third-party measurement partner.

Weak denotes testing discipline without a named design: A/B and split tests, variant language, stated statistical significance, confidence intervals, stated sample sizes.

Method is scored as applied rather than advertised. A method is counted only where an execution or inference verb appears adjacent to it. "We ran a matched market test across six DMAs" is counted; "Media Mix Modeling and Incrementality testing" in a services menu is not, and neither is "conversion (CRO + A/B testing)" in a capability list. Advertised capability is not evidence that a published number was measured. Imposing this requirement reduced the strong-tier count from 42 to 39 firms in the first run, and separates measurement of what firms do from measurement of what they sell.

2.5 Unit of disclosure

Disclosure is treated as a property of the page rather than of the sentence. A method stated anywhere on the page carrying a claim is scored as disclosed. This is the construction most favorable to the publishing firm, adopted deliberately: the observed effect is large enough that the most generous available standard costs nothing and forecloses the obvious objection.

The alternative construction, a character window around each claim, was rejected. The measured rate varies by an order of magnitude according to where the window is drawn:

Window	Share of claims with method language
140 characters	2.3%
400 characters	4.9%
1,000 characters	8.8%
2,500 characters	13.1%
Anywhere on the same page	23.0%

Median distance from a claim to the nearest method phrase, conditional on one existing, is 1,787 characters, or approximately a screen and a half of text. A statistic that moves tenfold across defensible analyst choices is not a suitable headline. The page-level ceiling is.

---

3. Results

3.1 Distribution of method language across the population

Of 369 participating firms:

	Firms	Share
Describe running a named design	51	13.8%
Testing language, no named design	62	16.8%
No comparison of any kind	256	69.4%

Classification rule. The classification is a site-level binary and the threshold is a single occurrence. A firm is placed in the first row if, anywhere across all of its retrieved pages, it describes having run a named design even once. It is placed in the second row if it never does that but describes testing discipline (an A/B or split test, a stated significance level) at least once. It falls into the third row only if neither appears anywhere on its site. The bar is therefore as low as it can be set while remaining a bar: one sentence, once, on any page. 69.4% of firms do not clear it.

Figure 1 gives the distribution. The proposition that the industry does not measure is not supported. Approximately three firms in ten show evidence of some comparison condition, and 13.8% describe a named design. The analytically interesting subpopulation is that 13.8%, because it is the group for which capability is not in question.

Computed on the 390 media-placing hosts prior to removal of the twenty-one non-participants, the same three rows read 13.1%, 16.4% and 70.5%. The population correction moves the headline share by 1.1 percentage points.

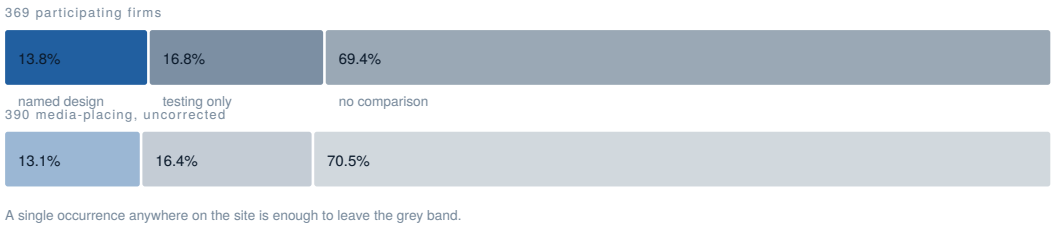

Figure 1. Method language across the 369 participating firms, as a share of firms. A firm enters the first band if its site describes running a named design even once, the second if it describes testing discipline but never a named design, and the third only if neither appears anywhere. The lower bar repeats the measure on the 390 media-placing hosts before the twenty-one non-participants were removed, and shows that the population correction moves the headline share by roughly one point.

3.2 Principal result

Restrict the population to firms whose possession of the methodology is not in question, on the evidence that their own websites describe having run a named experimental design. There are 51 such firms. Across 4,820 of their published numeric claims, 76.6% appear on a page that does not mention a method in any form.

The phrase carries its full weight. Not a named design, not a control or holdout group,

not an exposed and unexposed cell, not a matched market, not an A/B or split test, not a stated significance level or confidence interval, not a sample size, not a named measurement partner, not a period over which a comparison was drawn, and not a statement that no comparison was made. The generous construction of section 2.5 applies throughout: any one of those appearing anywhere on the page, adjacent to the claim or a thousand words away from it, counts as disclosure. Three quarters of the claims published by firms that demonstrably run designs clear none of it.

At page level, 77.8% of these firms' 1,057 case study pages carry no method. Widening the qualifying threshold from a named design to any applied method (113 firms and 9,045 claims) gives 78.3%, and at page level 81.2%. The result is not sensitive to where the threshold is set, and moves in the wrong direction for a capability explanation: the firms with the most demonstrated method are not the ones attaching it to their numbers.

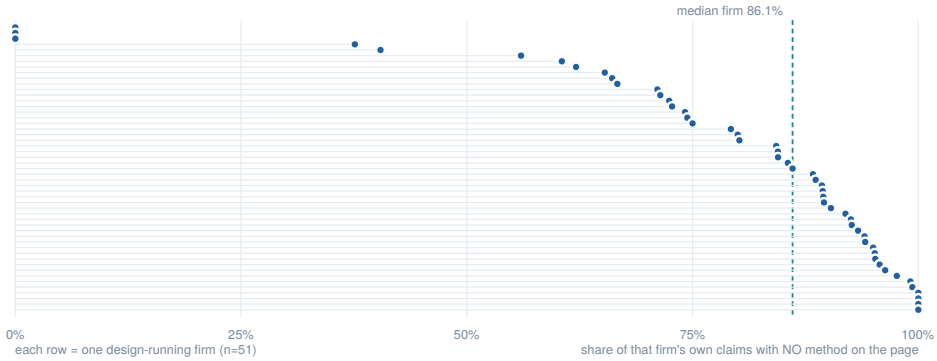

Figure 2. Non-disclosure rate for each of the 51 firms that describe running a named design, one point per firm, ordered ascending. The horizontal position is the share of that firm's own published claims sitting on a page with no method. Capability is held constant across every point, so the spread is variation in publishing practice alone. The median firm is at 86.1%, above the claim-weighted pooled figure of 76.6%, because the firms publishing the most claims disclose slightly more often than the median firm does. Four firms at the far left disclose on every claim, which establishes that the behavior is available to any of them.

Figure 2 gives the same result firm by firm, and shows that it is not produced by a handful of large publishers: the mass sits against the right-hand edge across the whole group.

What this rules out. The natural first explanation for undisclosed results is that the firm did not measure. That explanation is unavailable here by construction: every firm in this comparison published, elsewhere on its own site, a description of a design it ran. The second explanation is client confidentiality. It is also unavailable: these are marketing pages the firm chose to publish, and the method is not the confidential part. The client's data is, and the client is frequently named while the method is not. What remains is that stating the method is not part of the form.

Replication. An earlier crawl of 410 websites and 255 media-placing firms, with 39 firms in the design-running group, gave 77.2%. The corpus subsequently grew by 53% and

the qualifying group by 12 firms, and the estimate moved by six tenths of a percentage point.

	First crawl	Second crawl
Media-placing firms	255	390
Design-running firms	39	51
Their claims with no method on the page	77.2%	76.6%
Widened to any applied method	78.8%	78.3%

Robustness to the population definition. All 51 design-running firms fall inside the participating population, so the principal estimate is arithmetically identical before and after the population was redefined:

	Firms	No comparison of any kind	Design-running	Their claims with no method
Media-placing, uncorrected	390	70.5%	51	76.6%
Participating firms	369	69.4%	51	76.6%
...classified as agencies	271	69.7%	38	74.5%
...classified as platforms and vendors	98	68.4%	13	80.5%

The disaggregated comparison. Agencies and platforms are indistinguishable. 13.8% of participating firms describe running a named design; disaggregated, 14.0% of agencies and 13.3% of platforms. 69.4% publish no comparison of any kind; disaggregated, 69.7% and 68.4%. Agreement to within a percentage point on two independent measures is not the pattern expected of two substantively different populations, and is consistent with a classifier separating firms whose publication behavior is the same.

The apparent exception does not survive correction for clustering. The within-firm rate reads 74.5% for agencies against 80.5% for platforms, and pooling across all claims yields $z = 4.6$ for that difference. Claims are not independent observations: they cluster within firms, and a small number of large platform sites contribute thousands each. Treating each firm as one observation, the agency mean is 77.5% ($n = 38$, sd 27.3) and the vendor mean 80.1% ($n = 13$, sd 14.9), giving a Welch t of 0.42. The pooled firm-level mean is 78.2%, 95% CI $\pm$6.7 percentage points, an interval containing the claim-weighted estimate of 76.6%.

The phenomenon is therefore a property of how advertising results are published by the firms that produce them, and is not specific to advertising agencies.

Identification. The within-firm comparison is the estimate to weight most heavily, because each firm serves as its own control. Some publication threshold certainly governs which results a firm elects to disclose. Whatever that threshold is, it operates identically on the pages that disclose method and those that do not. The principal objection to any study of published claims, that only self-selected results are observed, cancels within firms.

3.3 A rejected hypothesis

The preregistered expectation was that weaker methodological language would accompany larger claimed effects. It is not supported.

Percent-unit claims from the 51 design-running firms:

	n	Median	IQR	p90
Page discloses a method	721	51.0%	20 to 96	220
Page does not	2,493	50.0%	22 to 99	283

The difference is one percentage point. Within firms, among the 29 with at least five claims on each side, disclosing pages carry the lower median claim in 11 of 29, below rather than at chance. Across firms in the earlier crawl the ordering was non-monotone: 48% for the design-running group, 66% for the middle tier, 54% for firms showing no comparison. A dispersion-based specification also fails.

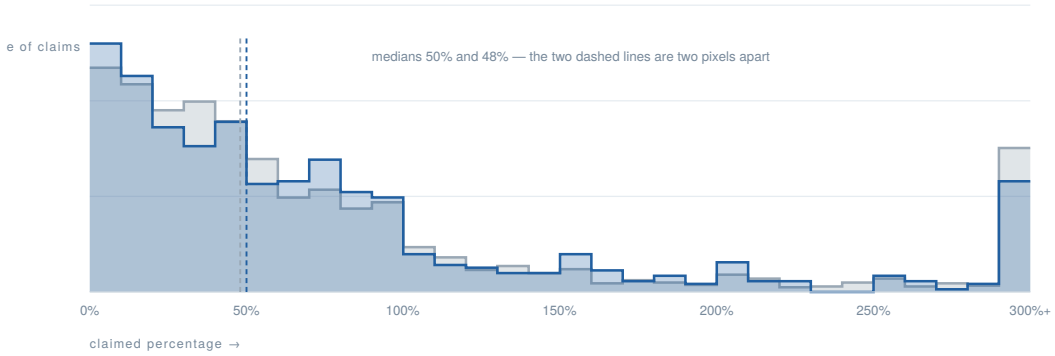

Figure 3. Distribution of claimed percentages published by the 51 design-running firms, split by whether the page states a method. Navy is the disclosing set, grey the non-disclosing one; both are shares of their own group, so the two curves are directly comparable despite unequal sample sizes. The axis is truncated at 300% for legibility. The two distributions have the same shape and their medians differ by a single percentage point, which is the null stated visually: a reader shown only the number cannot infer whether a method sits behind it.

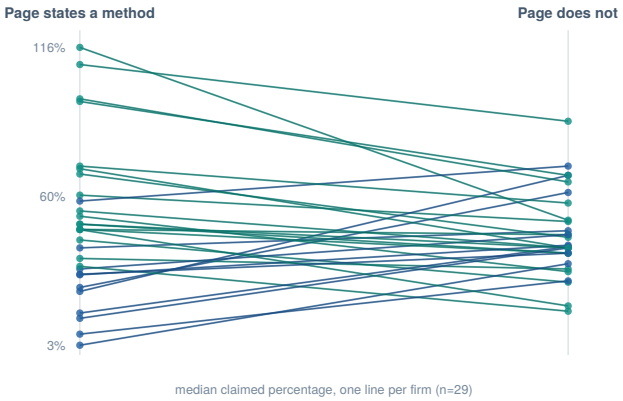

Figure 4. The same comparison made inside each firm. Each line is one of the 29 design-running firms publishing at least five percent-valued claims on both sides, joining the median claim it publishes with a method to the median it publishes without one. The hypothesis predicts that every line tilts upward to the right. Navy marks the 11 firms where it does; teal the 18 where it tilts the other way. The dense crossing through the middle of the range is the result: within a firm, the presence of a method carries no relationship to the size of the number published.

Figure 3 shows the two distributions overlaid and Figure 4 makes the same comparison inside each firm. The null is reported because it is the more consequential result. Were large claims reliably the undisclosed ones, a reader could apply a protective heuristic and discount large numbers. The null establishes that no such heuristic is available. Claim magnitude carries no information about whether the claim was measured, leaving disclosure itself as the only available signal. Disclosure is absent from more than three quarters of the claims published by firms that demonstrably run designs.

3.4 Interpretive properties absent from published claims

Four properties are required in order to interpret a published percentage. Each was measured on the 4,252 claims drawn from pages where document structure was preserved, scoped to the claim's own block and its immediate neighbors.

Each rate is reported twice: as measured by the automated classifier, and corrected for that classifier's measured error in both directions. The corrected figure is the estimate to read. It is lower in three of the four cases, because the classifier misses genuine disclosures more often than it invents them, and a study whose principal finding is an absence must correct in the direction that works against its own conclusion.

Property	Absent, corrected	95% CI	Absent, raw	Kappa	Precision
Source of the figure	94.7%	82.7 to 98.5	99.9%	0.84	33%
Period over which achieved	84.9%	71.0 to 92.2	96.4%	0.65	57%
Concurrent activity	82.4%	67.0 to 91.4	99.2%	0.66	76%

Property	Absent, corrected	95% CI	Absent, raw	Kappa	Precision
Numeric baseline	99.2%	90.6 to 99.2	98.2%	0.54	42%

Reading the range. The spread across those four figures is the difference between the properties rather than an uncertainty band. A numeric baseline is the property most often missing, at 99.2%; enumerated concurrent activity is the property most often present, at 82.4% missing. The 95% confidence interval attached to each row is the uncertainty on that row. The intervals are wide because the correction extrapolates from a random sample of approximately forty rules-negative claims per property, and they are reported rather than smoothed. At the optimistic bound of every interval, three of the four properties remain absent from more than two thirds of published claims.

Absence across subgroups. The uniformity is the stronger result. A reader might reasonably expect the interpretive properties to be present somewhere: in a regulated vertical, perhaps, or among the firms that run designs. They are not. Evaluating the four criteria block-scoped across 4,138 claims from 88 firms where document structure was preserved:

Grouping	n	All four absent
Destination and tourism	251	99.2%
Ecommerce and DTC	326	95.7%
Healthcare	179	95.5%
Nonprofit and public	529	94.3%
Higher education	640	93.9%
Credit union and banking	289	92.7%
B2B tech	531	91.5%
Multi-location	298	90.9%
Firms describing a named design	858	97.1%
Firms describing testing only	1,255	93.4%
Firms describing no comparison	2,025	92.8%
Pages that disclose a method	392	96.9%
Pages that do not	3,746	93.6%

Every cell falls between 91% and 99%. Regulated verticals do not differ from unregulated ones. Firms that run designs do not differ from firms that do not, and if anything

sit slightly worse, which is a page-genre effect: a firm describing a design writes a longer page, and the claim's own block is a smaller share of it. Pages that disclose a method are no better on the other four properties than pages that do not, which establishes that method disclosure and interpretive completeness are separate behaviors, not two symptoms of one underlying diligence.

There is no subgroup in this corpus in which a reader is adequately served. The absence is the convention, not a failing concentrated among the least capable.

What an incomplete claim supports. A published percentage lacking all four properties is not a weak claim. The reader cannot determine what the figure is a proportion of, over what interval it accumulated, what else was running that could have produced it, or who computed it. There is no operation the reader can perform on that number: it cannot be validated, and it cannot be discarded on any principled ground either. That is what makes it corrosive rather than merely uninformative. It is indistinguishable on the page from a number that was rigorously produced, and section 3.3 establishes that its magnitude carries no signal that would separate the two.

The cost of the four conditions. Four of the five conditions set out in section 6 cost only a sentence. Stating what a percentage is calculated from, over what period, alongside what other activity, and from what source requires no experiment, no additional budget and no capability the firm does not already hold. Only the comparison condition requires additional work. The costless conditions are absent from the large majority of published case studies, uniformly across every dimension measured, which is the evidence on which the gap is characterized as one of convention rather than capability, and is the central claim of this paper.

The source property warrants a note. It appears once in 4,252 claim scopes. That figure was not accepted at face value and was tested with a deliberately over-firing net covering GA4, attribution, pixel, dashboard, tracking and verification language. The net returned 3.4%, of which the majority were instances of "attributed to" used as a causal assertion and not as a disclosure of source. The near-total absence is a real feature of the corpus.

3.5 What predicts disclosure

The results above are prevalence. This section asks what disclosure covaries with, both to test the convention account against alternatives and to establish where in the market the absence is concentrated.

3.5.1 The distribution is not a spread, it is a mass at zero Among the 223 firms publishing at least three case study pages, the median firm discloses a method on 0.0% of them. 145 firms, 65.0%, disclose on none. The 90th percentile firm discloses on 25.0%. Four firms disclose on all of their pages.

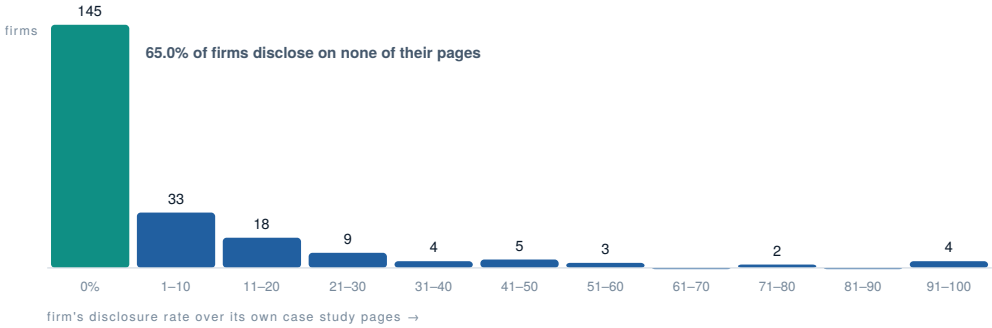

Figure 5. How the pooled non-disclosure rate is actually composed. Each bar counts firms by the share of their own case study pages that state a method, across the 223 firms publishing at least three such pages. The distribution has no centre: 145 firms, 65.0% of them, sit in the leftmost bar and disclose on none of their pages, and the median firm is at 0.0%. A pooled rate could have described an industry in which most firms disclose occasionally. It describes one in which most firms never do and a small number sometimes do.

Figure 5 shows the shape, and the shape matters for how the headline is read. A pooled rate of 76.6% could describe an industry in which most firms disclose sometimes. It does not. It describes an industry in which most firms never disclose and a small number sometimes do.

3.5.2 Disclosure rises with publishing volume, and the effect is partly mechanical

Case study pages published	Firms	Pages	Pages disclosing a method
1 to 2	44	64	1.6%
3 to 5	37	139	2.2%
6 to 14	82	753	7.4%
15 or more	104	2,430	11.3%

The gradient is monotone and substantial, and it is not wholly a behavioral finding: a firm publishing two case studies has less surface on which a method could appear, so detection opportunity rises with volume by construction. The correct reading is that the ceiling rises with volume and even at the top of the range the rate is 11.3%. Among firms with fifteen or more claim-bearing pages, 41% still show no comparison condition anywhere on their site.

3.5.3 Capability scales with firm size; disclosure does not On the 98 firms for which a headcount was obtained:

Headcount	Firms	Ever describe a named design	Case study pages disclosing a method
1 to 10	10	10.0%	13.0%
11 to 50	40	17.5%	12.7%
51 to 200	35	28.6%	4.3%
201 or more	13	38.5%	30.1%

Reading the two columns. Figure 6 plots both. The left is monotone and the right is not. The probability that a firm has ever run a named design rises steadily with headcount, from one in ten at the smallest firms to nearly two in five at the largest. A fixed-cost account of measurement capability predicts exactly that, since a measurement function does not scale down. The probability that a given published page states a method shows no corresponding gradient.

Figure 6. Capability and disclosure plotted against firm size, on the 98 firms for which a headcount was obtained. Left, the share of firms in each band whose site has ever described running a named design. Right, the share of those same firms' case study pages that state a method. Both panels are percentages on the same scale and are read independently; this is two charts, not one chart with two axes. The left panel rises monotonically with size, which is what a fixed-cost account of measurement capability predicts. The right panel does not, which is the dissociation: if disclosure followed from capability the two panels would have the same shape.

This is the dissociation the paper turns on, arriving from an independent direction. Capability is distributed the way capability is normally distributed: unevenly, by size, as a function of what a firm can afford. Disclosure is not distributed that way at all. If disclosure were a downstream consequence of capability, the two columns would track. They do not.

The headcount sample is 98 of 369 firms and the largest band holds 13. The gradient in the left column is the finding; the right column is reported as showing no gradient rather than as an estimate of one.

3.5.4 Disclosure varies by claim type

Claim type	Claims	Page discloses a method
Awareness or brand comparison	139	46.0%
Incrementality asserted without method	82	20.7%
Platform-attributed	512	17.0%
Multi-intervention	2,150	13.4%
Causal claim without method	751	13.3%
Benchmark-relative	230	10.4%
Bare before-and-after	7,404	9.6%
Input metric	1,534	8.9%
Year-over-year without control	1,051	8.7%

Awareness claims disclose at nearly five times the rate of before-and-after claims. Nowhere else in this study does the form of the claim predict its treatment. The explanation is instrumental rather than cultural: brand and awareness outcomes cannot be observed in a platform dashboard and require a survey instrument, so the method arrives attached to the number because there is no way to produce the number without commissioning it. Where the measurement apparatus is external and purchased, it gets named. Where it is internal and free, it does not.

The largest category by a wide margin is the bare before-and-after, at 7,404 claims and 9.6% disclosure. It is the industry's default published object.

3.5.5 Vertical differences do not survive within-firm comparison Page-level disclosure rates differ substantially across client verticals:

Vertical	Pages	Firms	Disclosing	95% CI
CPG and retail	125	37	30.4%	23.0 to 38.9
Multi-location	216	55	19.4%	14.7 to 25.2
Ecommerce and DTC	319	65	16.9%	13.2 to 21.4
Credit union and banking	118	43	11.9%	7.2 to 18.9
Destination and tourism	358	74	9.5%	6.9 to 13.0
Healthcare	207	64	8.2%	5.2 to 12.8
B2B tech	550	89	7.8%	5.9 to 10.4
Higher education	253	61	4.3%	2.4 to 7.6
Nonprofit and public	195	42	2.6%	1.1 to 5.9

A twelve-fold spread between the top and bottom of that table invites an obvious account: some client industries demand rigor and their agencies supply it, while others

do not and their agencies are permitted to publish without method. The account is plausible, it is commercially useful, and the data does not support it.

Figure 7 sets the raw rates beside the within-firm deviations. Firms serve multiple verticals, which makes the test available. 144 of the 252 firms publishing labeled case studies publish in two or more verticals, contributing 1,882 pages. Holding the publishing firm constant and measuring each vertical's deviation from that firm's own average disclosure rate:

Vertical	Pages	Raw rate	Deviation from the same firm's own average
Healthcare	113	8.2%	+3.3 pp (−1.4 to +8.1)
Credit union and banking	53	11.9%	+1.8 pp (−6.8 to +10.4)
Multi-location	186	19.4%	+1.5 pp (−2.5 to +5.4)
Ecommerce and DTC	299	16.9%	+1.0 pp (−1.6 to +3.6)
Destination and tourism	213	9.5%	+0.1 pp (−3.1 to +3.4)
B2B tech	457	7.8%	−0.2 pp (−2.2 to +1.9)
Nonprofit and public	136	2.6%	−0.3 pp (−3.1 to +2.5)
Higher education	134	4.3%	−1.5 pp (−4.8 to +1.7)
CPG and retail	102	30.4%	−1.7 pp (−6.5 to +3.2)
Legal and professional	50	7.3%	−2.8 pp (−8.6 to +3.0)
Real estate	31	6.5%	−8.3 pp (−12.9 to −3.6)

Result of the within-firm comparison. The ordering does not survive. CPG and retail has the highest raw disclosure rate in the corpus at 30.4% and a *negative* within-firm deviation: a firm publishing CPG work discloses slightly less on that work than on its own work elsewhere. Every deviation except real estate, on 31 pages, is indistinguishable from zero. A permutation test that shuffles vertical labels within each firm returns $p = 0.093$ for the observed spread.

The variance decomposition states it directly. Of the total page-level variance in disclosure, between-firm differences account for 51.9% and between-vertical differences for 4.8%.

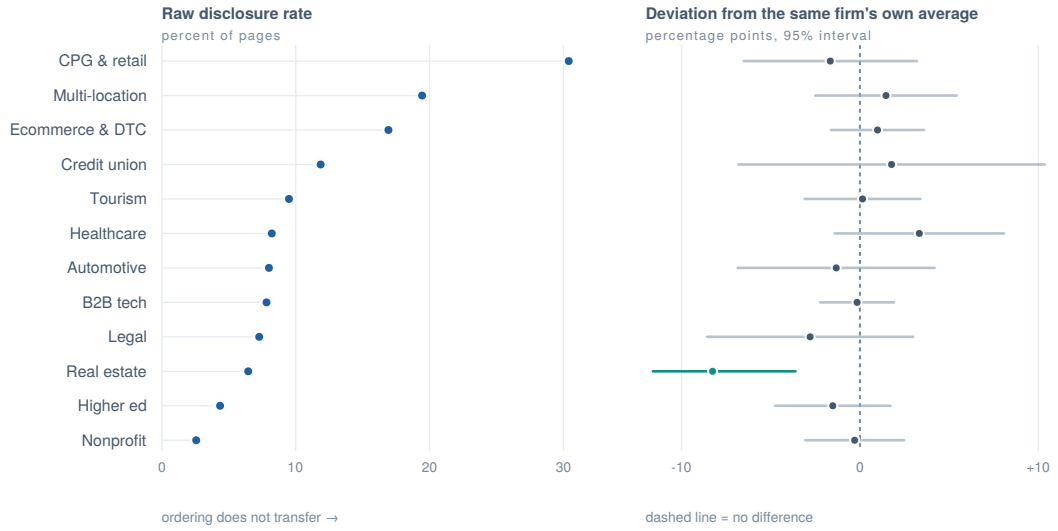

Figure 7. Whether client vertical explains disclosure. Left, the raw page-level disclosure rate for each vertical, a twelve-fold spread that appears to show some client industries obliging their agencies to report more carefully. Right, the same verticals measured as deviation from the publishing firm's own average, across the 144 firms publishing in two or more verticals, with 95% intervals; the dashed line is no difference from that firm's own behavior. Rows are in the same order in both panels, and the units differ, which is why these are paired panels and not a slope chart. The left-hand ordering does not reappear on the right: CPG carries the highest raw rate and a negative deviation, and every interval except real estate, on 31 pages, crosses zero. The verticals differ because different firms publish in them.

Interpretation. The vertical spread is a firm-composition effect. Verticals do not differ because clients in them demand different things. They differ because different firms publish in them, and disclosure is a property those firms carry across everything they publish. CPG work looks rigorous because firms that disclose happen to publish CPG work, not because CPG buyers oblige their agencies to say more.

This is the same structure as the agency-versus-vendor result in section 3.2, and it strengthens the paper's central claim considerably. If disclosure were responsive to client demand, it would move inside a firm as that firm moved between a regulated depository client and an unregulated ecommerce one. It does not move. Two verticals in this table, credit unions and healthcare, carry substantive regulatory exposure around substantiated claims, and neither shifts its agency's behavior by a measurable amount. Disclosure is not a response to what the client requires. It is a house style, and it is set at the firm.

---

4. Inter-rater reliability

Each of the four properties was judged independently by two models at temperature zero, each required to quote the span relied upon, across every rules-positive claim and

a random sample of rules-negatives. Cohen's kappa is reported rather than raw agreement: at approximately 96% "no" prevalence, two raters that always answer no agree 96% of the time while demonstrating nothing.

Three of the four criteria are well defined; one is marginal. Source of the figure, at 0.84, is almost perfect agreement. Concurrent activity at 0.66 and period at 0.65 are substantial. Numeric baseline at 0.54 is moderate, and is the property independent readers apply least consistently. It is reported rather than dropped, and should be weighted accordingly.

Two criteria were rewritten during the study, in both cases driven by rater disagreement and not by the authors' judgment.

The first asked whether the text states what a number is a proportion of. Raters split on every positive instance, because "ad spend" is a truthful answer that leaves the claim exactly as unevaluable as saying nothing at all. Requiring a numeric baseline resolved the disagreement.

The second asked whether other marketing activities were named, a condition that abstract program language technically satisfies. Human review of the disagreements established that the operative rule was enumeration: "a full-funnel strategy" was rejected, "lapsed reactivation, acquisition, renewal" was accepted. Rewriting the criterion to require two or more separately identifiable activities moved kappa from 0.40 to 0.66 and precision from 55% to 76%.

Disagreement between capable readers is diagnostic of the criterion rather than of the readers. In neither case was a rater wrong; in both cases the question was.

---

5. Limitations

Accuracy is not observed. No claim is made here that any published number is false.

Published pages are observed, not internal reporting. Whether a firm discloses method to its client, or to the public through any channel other than its own case studies, is not measured and cannot be inferred from these data.

The frame is not representative. It is a regionally and vertically weighted corpus assembled by search and client-side enumeration.

Claims are selected by publishers. Section 3.2 sets out why the within-firm comparison is robust to this selection and why the pool-level rates are not.

Document structure was available on a subset. The four properties in section 3.4 are measured on 4,252 claims from pages crawled after the extractor preserved block boundaries, not on all of them.

Automated classification carries measured error. Rules precision ranges from 33% to

76% and is reported per property. False negatives, where the rules score a property absent and both raters score it present, imply that reported absence rates are overstated and that corrected values are lower.

The vertical labeler is rules-based and unvalidated. Vertical assignment in section 3.5.5 comes from term matching restricted to page title and URL slug, a setting chosen for precision over coverage; 27.5% of pages carry no label and are excluded from that analysis. The within-firm result is robust to label noise in a way the raw rates are not, because misassignment moves pages between verticals within the same firm and attenuates the estimated vertical effect toward zero. The finding is that the effect is zero, so a reader should treat the raw ordering as the less reliable of the two tables.

Firmographic coverage is partial. Headcount was obtained for 98 of 369 firms, and section 3.5.3 is reported on that subset. The largest headcount band holds 13 firms.

The population boundary is a judgment. "Firms that support client advertising outcomes" is operationalized through two automated properties: whether a site evidences paid media execution, and whether it falls under a named non-participant rule. The second was applied retroactively after the crawl. Both are inspectable, and Appendix B lists every removal, but neither is a validated classifier. A reader drawing the boundary differently would recover the pool-level rates of section 3.1 to within a point or two. The within-firm result in section 3.2 does not depend on the boundary, since every design-running firm falls inside it under all definitions tested.

---

6. Criteria applied by the authors

No standard is proposed for general adoption. The following are the conditions this firm requires of a reported number before standing behind it, published here on the principle that a paper concerning disclosure should disclose its own.

1. **A numeric baseline.** What the figure is a proportion of, in quantities a reader could check.
2. **A comparison condition.** What the result is measured against, or an explicit statement that no comparison was constructed.
3. **A stated period**, together with whether it was fixed before the result was observed.
4. **Concurrent activity, enumerated**, where several activities ran together.
5. **The source of the figure.** Platform-reported, firm-calculated, or independently produced.

Four of the five cost a sentence. The second is the one requiring additional work.

Appendix A. Instrument errors and corrections

Every instrument used in this study was initially wrong, in each case in a manner that would have changed the published estimates. They are recorded because a methods section reporting no errors is reporting that no one examined the instrument.

The prospecting gate selected on the outcome variable. A filter removing firms that already referenced a control was, in an earlier design, also filtering the analysis corpus, which guarantees the expected result independent of its truth. Prospecting and analysis now write to separate outputs.

Advertised capability was read as applied method. A services menu listing "Media Mix Modeling and Incrementality testing" scored a firm as running designs. Requiring an adjacent execution verb corrected this.

Named measurement vendors fired on citations. "According to Nielsen, streaming now accounts for 37.7% of total TV viewing" scored as a measurement partnership. Vendor names now require study or measurement context.

A trailing word boundary rendered two patterns unmatchable. Any alternative terminating in a character class could never match, and did so silently.

The vertical labeler dropped 30% of pages and had poor precision on the remainder. Permitting a single high-precision term to label a page reduced unlabeled pages from 30.0% to 10.4% at approximately one-third precision, because navigation lists every industry a firm serves and a single occurrence of "cannabis" in a menu labeled a healthcare case study as cannabis. Restricting the shortcut to the page title and URL slug gave near-perfect precision at 27.5% unlabeled. Term lists cannot deliver both.

Half of all claims could not be located in the structured text, because quotes were extracted from normalized text and span block boundaries. Two-way normalized matching resolved this.

Two of the four criteria were rewritten mid-study, in both cases driven by rater disagreement. See section 4.

The non-participant filter never ran against the analysis corpus. The rules excluding trade press, directories and associations were applied at discovery only. Sites seeded before those rules were written remained in the population indefinitely, because nothing re-tested them. This is the same class of error as the prospecting gate above, and the third instance in this project of a filter being improved without the data it had already admitted being re-filtered. Retroactive application of the current rules removed the twenty-one firms listed in Appendix B. Every filter in the pipeline now supports retroactive application.

Appendix B. Firms excluded as non-participants

Twenty-one of the 390 media-placing hosts participate in no client advertising outcome. They enter a domain crawl because they publish on the same subject matter. Each is removed under a named rule, and each rule was tested against the full media-placing set for false exclusions before being applied; the test returned none. Firms with an agency term in the domain are protected by an override, so that a firm named `impressivedigital` is not excluded for containing the string `press`.

Rule	Hosts
Business and trade press	investopedia.com, finextra.com, franchising.com, businessalabama.com, comstocksmag.com, appsruntheworld.com, citizenportal.ai, scitechnol.com, gulfshorebusiness.com, idahobusinessreview.com, oregonbusiness.com, vcstar.com
Research vendor	destinationanalysts.com, zartico.com, adara.com
Subsection of a larger organization	partner.expediagroup.com, blog.milestoneinternet.com, blog.marketingdatascience.ai
Industry press and events	phocuswire.com, hospitalitynet.org
Destination marketing organization	travelnevada.biz

None of the twenty-one is a design-running firm, which is why the principal estimate is unchanged by their removal.

Appendix C. Figure provenance

The seven figures are drawn from the census outputs by `figdata.mjs`, which emits the underlying series as JSON, and rendered by a standalone page held as a working artifact. Extraction to vector resolves every stylesheet variable to a computed value so each figure stands alone, and conversion to PDF preserves it as vector at any scale. Nothing in any figure is drawn by hand or adjusted after rendering.

Firm identities are withheld on every figure. Figures 2 and 7 plot one mark per firm and would otherwise reconstruct a named ranking of firms by their disclosure rate, which this paper's editorial line declines to publish.

Two properties of the underlying data are worth stating beside the figures. The dot plot and the histogram are the two that carry the argument visually: the first shows that the

firms possessing the methodology sit at the top of the non-disclosure range, and the second that the industry distribution is a mass at zero and not a spread. The paired-panel figure is the one that forecloses the reading a commercially motivated reader will otherwise reach for, which is why it is reported as two panels on their own scales instead of a single chart that would have required two y-axes.

Window sensitivity, reported as a table in section 2.5, is a candidate eighth figure and is defensible as a table.

Appendix D. Computational disclosure

This study used automated classification at two stages, and both are disclosed in full because the paper's central figures depend on them.

D.1 Rule-based classification

All extraction, eligibility screening, method-language detection, vertical labeling and non-participant exclusion were performed by deterministic regular expressions and procedural code in Node.js, with no model involvement. Every pattern is reproducible and the operative ones are printed in Appendix E.

D.2 Model adjudication of the interpretive criteria

Section 4's inter-rater reliability was obtained from two independently prompted commercial language models, each judging every rules-positive claim and a random sample of rules-negatives, with a human third rater adjudicating disagreements.

	Rater 1	Rater 2
Vendor	OpenAI	Google
Model, as pinned	`gpt-5.4-mini-2026-03-17`	`gemini-3.5-flash`
Endpoint	`/v1/chat/completions`	`generativelanguage.googleapis.com/v1beta`
Temperature	0	0
Response format	Forced JSON object	Forced JSON (`application/json`)

Model identifiers were pinned to an exact version in configuration, not to a rolling alias, so the run is reproducible against those versions for as long as the vendors serve them. Where the OpenAI endpoint rejected a `temperature` parameter, the call was retried without it and that fact recorded; no other parameter was varied.

The raters were blind to one another and neither was shown the rules-based label. Each was given the claim's own text scope, the numeric value in question, and one criterion phrased as a yes-or-no question, and each was required to return the exact span it relied upon:

```
{"answer": "yes" | "no",
 "span": "<the exact words you relied on, or empty string>",
 "confidence": "high" | "low"}
```

The prompt instructed the rater to answer only about the number given, and stated that text concerning a different nearby number does not count. Requiring a quoted span is the substantive control: a rater that cannot produce the words it relied on has not made a judgment that can be checked, and spans were used in the human adjudication of disagreements.

Agreement is not correctness. Two models can agree and both be wrong, and the paper does not treat concordance as validation. Cohen's kappa is reported as a property of the criterion, namely whether it is defined well enough for independent readers to apply consistently. The human third rater adjudicated every disagreement, and that adjudication produced the two criterion rewrites recorded in section 4.

Transient failures were retried with backoff. HTTP 429, 500, 502 and 503 responses are transient; recording them as judgments would leave gaps that look like data. Residual failures are counted and reported per rater, not silently dropped.

No model was used to write, select, or interpret any figure reported in this paper.

D.3 Data and code availability

The instrument and the derived data are available on request and are being prepared for public release: the crawler, classifier, criteria detectors and adjudication harness, together with the census outputs (`census-claims.jsonl`, `census-agencies.jsonl`) and the scripts that reproduce every figure and table (`census-current.mjs`, `relationships.mjs`, `criteria-dims.mjs`, `merged-population.mjs`, `figdata.mjs`). Raw crawled page text is not redistributed, as it is third-party copyrighted material; the derived claim records carry the source URL so any figure can be checked against the live page.

Requests go to trevor@darkwavemarketing.science. A request does not need a reason attached. This paper reports that four fifths of an industry publishes numbers a reader cannot check, and a request to check these ones is the response it asks for.

This document and its eight-page summary are published without registration at https://darkwavemarketing.science/notes/method-disclosure-in-case-studies.

Appendix E. Operational definitions

The patterns below are the operative ones. They are printed rather than described because a description of a regular expression is not a definition of it, and every prevalence figure in this paper is downstream of these.

E.1 Strong method language: a named design with a comparison group

```
control group | control cell | controlled (?:test|experiment|trial)
holdout | hold-out | held out | holdout group | holdout market
matched market | synthetic control | randomi[sz]ed | randomi[sz]ation
test and control | test vs\.? control | test cell | treatment group
exposed (?:group|cell|cohort) | unexposed | non-?exposed
geo test | geo-test | geo experiment | geographic holdout
lift study | lift test | brand lift study | conversion lift study
incrementality (?:test|study|experiment) | counterfactual
difference-in-difference | diff-in-diff
(?:kantar|disqo|dynata|upwave|nielsen|analytic partners|rockerbox|northbeam
  |incrmntal|adelaide)[^.]{0,60}(?:study|lift|test|panel|measur|survey|partner)
(?:study|lift|test|panel|measur|survey|conducted with)[^.]{0,60}(?:kantar|disqo
 |dynata|upwave|nielsen|analytic partners|rockerbox|northbeam|incrmntal|adelaide)
incremental(?:ity)?[^.]{0,60}(?:test|study|experiment|holdout|control|design|measur)
(?:test|study|experiment|holdout|control|design|measur)[^.]{0,60}incremental(?:ity)?
```

Case-insensitive, alternatives joined. The two bounded vendor-name patterns replaced bare vendor names after a hand read established that "According to Nielsen, streaming now accounts for 37.7% of total TV viewing" was scoring as a measurement partnership. The same bounding applies to "incremental," which counts as method language only where a method sits within sixty characters of it.

E.2 Weak method language: testing discipline without a named design

```
a/b test | a/b-test | \bab test | split test | multivariate test
variant [ab]\b | \bvariant\b[^.]{0,40}\b(?:won|winner|outperform|lift)
statistical(?:ly)? significan | significance:? \d | confidence interval
p-value | p < ?0 | sample size | baseline (?:period|campaign)
pre-?period | post-?period
```

E.3 Applied rather than advertised

A method in E.1 or E.2 is counted only where one of the following appears within 200 characters of it. This is the distinction between measuring what firms do and measuring what they sell.

```
\b(?:ran|run|conduct\w*|execut\w*|implement\w*|launch\w*|deploy\w*|perform\w*
  |measur\w*|validat\w*|reveal\w*|determin\w*|confirm\w*|indicat\w*|carried out
  |set up|showed|shown|found|proved|proven|compared (?:to|against|with)|versus
  |vs\.?|against a|resulted? in|over the course of
```

```
  |during the (?:test|experiment|study))\b
```

E.4 The four interpretive criteria

Each criterion is evaluated on the claim's own text block plus one block either side, and each runs its exclusion patterns *before* its positive pattern, so that a citation containing a number does not satisfy a criterion.

Numeric baseline.

```
\bfrom \$?\d[\d,.]*\s*(?:to|→|->|up to)\s*\$?\d | \bout of \d
| \bbaseline of \$?\d | \bstarting (?:from|at) \$?\d
| \bcompared (?:to|with) (?:a )?(?:baseline|prior|previous) (?:of )?\$?\d
| \bon a base of
```

Excluded by: the citation pattern (E.5), and a year-range pattern, since "the recession years of 2008 and 2020" is not a baseline.

Period.

```
\b(?:over|during|within|across|in|after) (?:the )?(?:first |past |last )?
  \d+[- ]?(?:day|week|month|quarter|year)s?\b
| \bin \d+ days\b | \bQ[1-4] (?:19|20)\d\d\b
| \b(?:19|20)\d\d\s*(?:to|-|–)\s*(?:19|20)\d\d\b
| \b\d+[- ]?(?:day|week|month)s? (?:campaign|flight|test|period|window)\b
```

Excluded by: the citation pattern, and a company-tenure pattern, since firm age and relationship length are not measurement windows.

Concurrent activity, enumerated.

```
\b(?:alongside|in addition to|combined with|as part of|together with
  |while (?:also|simultaneously)|paired with|coupled with)\b
```

Additionally requires a named marketing channel in the same clause, and excludes a not-an-intervention pattern covering "worked alongside ", "alongside " and "ads displayed alongside content". This criterion was rewritten mid-study from *named* to *enumerated* on rater disagreement; the rewrite moved kappa from 0.40 to 0.66 and precision from 55% to 76%.

Source of the figure.

```
\b(?:as measured (?:by|in)|measured (?:by|using|via) (?:a |the )?[a-z]
  |platform[- ]reported|reported (?:by|in) (?:google|meta|facebook|ga4|adobe|the platform)
  |attribution window|per (?:google|meta|facebook|ga4|adobe) (?:analytics|ads)
  |third[- ]party (?:measurement|verified)|verified by)
```

E.5 Exclusion patterns

Third-party citation, applied before every criterion:

```
\b(?:according to|as reported by|per a |a \d{4} (?:study|survey|report)
  |research (?:by|from|shows|found)|study of \d|survey of \d|\bsource:|\(source\b
  |statista|nielsen report|pew ?research|gartner|forrester|mckinsey|deloitte
  |hubspot (?:reports|found)|\bstudy after study
  |\b(?:patients|consumers|shoppers|buyers|users|donors|travelers|marketers
    |respondents) (?:have|are|say|report|prefer|expect)\b|\bsimilarly,)
```

Company tenure, applied to the period criterion:

```
\b(?:years? (?:ago|in business|of experience|of partnership)
  |since \d{4},? (?:we|our|the (?:agency|firm|company))
  |for (?:over |more than |nearly )?\d+\+? years?,? (?:now|we|our|the|[A-Z])
  |\b(?:over|more than) \d+ years? of\b|founded in|established in
  |celebrating \d+ years)
```

This pattern carries no trailing word boundary, deliberately: an alternative ending in a character class can never satisfy one, and the defect that taught us this is recorded in Appendix A.

E.6 Claim location

Claim quotes were extracted from normalized text in which block breaks became spaces, so a quote routinely spans two or three blocks and a naive substring search finds nothing; a first pass lost 49% of claims this way. Matching is therefore two-way on normalized text: the block may contain the start of the quote, or the quote may contain the whole block, as occurs when a quote spans several.