Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

What asserted weights are worth: pricing a receptivity index in households and pre-specifying its test

Nicholas Sokol

Nsokol@utk.edu

University of Tennessee at Knoxville

Research Article

Keywords: composite indicators, receptivity, weight learning, hierarchical Bayes, geo-randomized experiments, pre-specification

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11543865/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: Competing interest reported. The author develops commercial market-analysis software that implements a version of the GCEIA index evaluated in this paper.

Abstract

Agencies and firms allocate by rankings, so a specification error becomes a targeting error. We audit GCEIA, a deployed five-dimension composite of 3,106 U.S. counties. On the pre-specified proxies (2020 Census self-response and turnout, 3,097 counties), fitted models beat the composite (gradient boosting ρ = 0.758 and 0.770 against 0.634 and 0.157). Two dimensions do not earn their weight on one proxy, and the proxies disagree in sign about two. We revise the index on theory, not on those outcomes, and freeze it (GCEIA-R). The revision improves the Census correlation (ρ = 0.666, difference + 0.032, 95% CI [0.018, + 0.047]) and is neutral on turnout. On turnout the asserted and fitted top deciles share only 30 of 310 counties. Priced in households, rank disagreement overstates the cost: a plan contacting the top tenth of households reaches 73.9% self-responders under the asserted index against 67.7% untargeted, and neither the fitted model nor GCEIA-R differs detectably (at the one-tenth budget). The decisive test has not been run. We pre-specify it as a protocol: a five-year geo-randomized design, hierarchical weight learning from the GCEIA-R prior, and a top-decile capture bar, sized by design-based power analysis. Its contribution is the frozen design, not a result.

1. Introduction

Every composite score that ranks places is, in use, an allocation rule. A county near the top of a receptivity ranking gets the outreach visit, the campaign spend, or a place on the site shortlist. A county below the cut gets none of them. A mis-specified index therefore does more than describe places wrongly. It sends real effort to the wrong places and withholds it from the right ones, at full price, and it does so silently because the score arrives without the audit trail that would expose the mistake. We measure that cost for one deployed index. We ask which of its dimensions earn their weight against outcomes and which are carried by the others, and we ask what a revision disciplined by theory recovers. Our main contribution is the audit: a deployed index held to outcomes, a revision frozen on theory alone, and the disagreement between asserted and learned rankings priced in households. We close by pre-specifying the test, and the decision layer built on it, that would settle whether any version of the score deserves the allocations it guides.

By receptivity we mean the propensity of a place's residents to respond when they are contacted, whether by a campaign, an agency, or a firm. It is a behavior and not an attitude, and our central finding is that it is not one behavior. The index we study is GCEIA, named for its five dimensions: Geography, Culture, Economics, Identity, and Accessibility. It scores 3,106 continental U.S. counties from eight American Community Survey (ACS) inputs. Box 1 lists the dimensions, their inputs and weights, and what the revision changes.

Box 1. GCEIA and GCEIA-R at a glance. Each input is converted to a national percentile, the five dimension scores are standardized across counties, and the composite is their weighted sum.

Dimension	Construct	GCEIA input (ACS)	Change in GCEIA-R	Weight
Geography	Settlement scale and concentration	Percentile of log population, with a constant	Density replaces population	0.15
Culture	Education	Bachelor's-or-higher rate	Unchanged	0.20
Economics	Income and its inequality	Median household income, and Gini index (inverted)	Unchanged	0.30
Identity	Age and household structure	Median age and average household size (both inverted)	Unchanged	0.15
Accessibility	Connectivity	Broadband rate, and mean commute time (inverted)	Commute input dropped	0.20

A composite index is a claim disguised as arithmetic. Its weights assert that some dimensions of a place matter more than others for the behavior the index is meant to anticipate, and that assertion is rarely tested against the behavior. Reviews of composite-indicator practice (Greco et al., 2019; Paruolo et al., 2013; Saisana et al., 2005) document the pattern. Weights are justified by expert judgment, budget allocation, or principal components, and robustness is assessed by perturbing the weights, not by confronting the ranking with an outcome. The sensitivity analysis of the social vulnerability index by Schmidtlein et al. (2008) is the type specimen. When the outcome check is finally run, the results are uncomfortable. In a companion study of the index used here (Author, 2026a), a static five-dimension composite of county receptivity tied equal weights almost exactly ($\rho = 0.987$), was not superior to all of five simple comparators on either of two public response proxies, and lost outright to a single variable, the bachelor's-degree rate, on one of them.

The setting is not peculiar to this index. Composite measures are the working instruments of applied spatial analysis. Deprivation scores built from census inputs steer regeneration resources and area-based interventions (Norman et al., 2024). Livability indices documented transparently, with data and code, quantify where densification policy is succeeding (Magnuszewski et al., 2025). Walkability and resilience scores steer pedestrian investment and flood planning (Cao et al., 2026; Bottoms et al., 2024). Such scores are built far more often than they are validated. Where validation occurs it is usually internal: sensitivity of the ranking to the weighting scheme (Libório et al., 2023; Punzo et al., 2022), propagation of sampling error (Feldmeyer et al., 2026), or agreement across scales and specifications (Hinojos and Grady, 2026). Behavioral validation against pre-specified outcome tests, with null results reported where the tests produce them, is rare in this literature. Until recently that included the literature of the index we study. We supply that missing half for one instrument, in the open, and we specify the test that could complete it. Table 1 sets this paper against the closest recent composite-index studies on the features that define its claim: whether the index was confronted with an outcome beyond its own construction, whether the test was specified in advance, whether unfavorable results were reported, and whether the weights were ever learned rather than asserted.

Table 1
This paper against the closest composite-index studies. "Not reported" means the feature does not appear in the study's report.

Study	Index and domain	Validation against held-out outcomes?	Pre-specified?	Null results reported?	Weights learned or asserted?
Bottoms et al. (2024)	Flood Resilience Assessment Index, U.S. census tracts	No. Convergent comparison with leading vulnerability indices	Not reported	Not reported	Asserted (mixed-methods framework)
Cao et al. (2026)	Walkability scores (TWS, PEWS), Nantes, France	Behavioral, not held out. Validated against observed pedestrian counts in the same case	Not reported	Not reported	Data-derived (entropy weights), not learned from outcomes
Libório et al. (2023)	Multidimensional composite, sub-indicator weighting schemes	No. The outcome studied is the composite's own spatial dependence	Not reported	Not reported	Asserted schemes, compared
Punzo et al. (2022)	SusTour-Index, sustainable tourism, Italian regions	No. Internal validation across 23 weighting and aggregation models, with uncertainty analysis	Not reported	Not reported	Asserted variants, model selected by uncertainty analysis
Feldmeyer et al. (2026)	Composite indices generally, input sampling error	No. Uncertainty propagation and sensitivity analysis	Not reported	Not reported	Asserted, treated as design parameters
This paper	**GCEIA and GCEIA-R, county receptivity**	**Pre-specified public proxies (now spent). The decisive held-out test is pre-specified in** Section 5 **and not yet run**	**Outcome tests pre-specified in the companion study. Decisive test pre-specified here**	**Yes. The turnout null is reported in full**	**Asserted, relabeled a prior. Outcome learning specified in** Section 5.3, **not yet estimated**

The rest of the paper takes the next step that such a result demands, in an order designed to keep the steps honest. First, we diagnose which dimensions of the index earn their weight against outcomes and which are carried by the others (Section 2). The diagnosis uses two independent instruments, fitted

models allowed to beat the index and one-dimension-at-a-time ablations, applied to the companion study's two pre-specified public proxies. Second, we revise the index on theoretical grounds and freeze it (Section 3). The revision imports the one repair the companion paper validated, density in place of population. It drops an input whose construct is wrong. It declines the changes that the diagnosis makes tempting but cannot justify. Third, we state the evaluation record of the revision on the same proxies, labeled for what it is, an evaluation on data that informed the diagnosis, and we price the disagreement between rankings in households reached (Section 4). Fourth, we pre-specify the test that can decide the question: a geo-randomized campaign design with a logging schema, a hierarchical model that learns weights outward from the revised index as its prior, and a pass bar on top-decile capture against the same free baselines, all fixed before any campaign data exist (Section 5). The design ends where the score is used. A decision layer (Section 5.6) converts learned receptivity into enter, avoid, or scale calls, priced against the rivals competing for the same counties. That is the form in which a receptivity index, if it earns the right, actually guides the targeting decisions of agencies and firms.

Two commitments govern the reporting, as in the companion paper. We report results whether they favor the index or not. And we call no version of the index superior to simpler alternatives unless a head-to-head test on a stated criterion, fixed in advance, says so. On the evidence available here, none does.

2. Diagnosis

The diagnosis is a case study, but the defects it documents are not idiosyncrasies of one index. They recur wherever composite indices are built once and then deployed as if settled, and they form a small taxonomy of failure modes. The first is *construct substitution*: an available input measures a convenient neighboring quantity rather than the construct named, as when population size stands in for settlement density or commute time stands in for access. The second is *weight persistence*: asserted weights are carried forward untested, so dimensions keep shares of the score that outcome data would not grant them. The third is *outcome-dependence of dimension value*: a dimension that helps one behavior hurts another, so any single weighting silently presupposes one behavior while claiming generality. The fourth is *reference dependence*: scores computed against a shifting batch of peers are not comparable across releases, scales, or time. GCEIA exhibits all four. Sections 2.2 and 2.3 document the second and third, and Section 3 repairs the first and fourth.

2.1 The index, the proxies, and the protocol

GCEIA scores 3,106 continental U.S. counties with complete data on eight ACS inputs, grouped into five dimensions (Author, 2026a). Geography combines the percentile of log population with a constant. Culture is the percentile of the bachelor's-or-higher rate. Economics blends the percentiles of median household income and (inverted) Gini. Identity blends (inverted) percentiles of median age and average household size. Accessibility blends the broadband-rate percentile with the (inverted) mean-commute percentile. We z-score the dimensions across counties and combine them with asserted weights W = [0.15, 0.20, 0.30, 0.15, 0.20]. The construction is deterministic and reproducible from public data.

The two outcome proxies are the companion study's, pre-specified there before the outcome values were seen. County turnout is total votes cast in the 2020 presidential election divided by ACS population. It is a population-wide participation rate with a known leak: turnout rises steeply with age, and two of the index's Identity inputs are age and household size. The 2020 Census self-response rate (final cumulative rate, U.S. Census Bureau, 2021) is a household's reply to an official mailed contact. It is the closer behavioral analogue to campaign response, and its denominator is housing units rather than residents, which reduces the age leak without removing it. Neither proxy is a marketing outcome. They are what public data offers, and we label every claim below with the proxy it was measured on.

The protocol is the companion study's Study 2 protocol, unchanged. We use Spearman ρ against the outcome. We use a state-clustered bootstrap (2,000 replications, states resampled) for intervals and for paired differences, and Holm correction across comparators. We call an index superior to the comparator set only if it beats every comparator with corrected intervals excluding zero. We compare fitted models out-of-sample under state-blocked ten-fold cross-validation, so that a model cannot win by memorizing a state's counties.

2.2 Fitted models beat the asserted composite on both proxies

We compared four models out-of-sample: the static index (F0), a global linear model on the five dimensions (F1), a model with Census-division coefficients shrunk toward the global model (F2), and gradient boosting on the eight raw inputs (F3). On turnout the asserted index scores ρ = 0.157, and the fitted models reach 0.564 (F1), 0.609 (F2), and 0.770 (F3). On Census self-response the index scores 0.634, and the fitted models reach 0.685, 0.737, and 0.758. The pattern is the same on both proxies and it is not subtle. Nearly all of the predictive signal available in these inputs is available only to a model allowed to learn its weights from an outcome. The global fitted weights also contradict the asserted ones. On turnout the largest fitted coefficient falls on Culture and the Geography and Identity coefficients are negative, while the asserted weights give Economics the largest share. An asserted weighting is not an approximation to the fitted one. On these proxies it is a different theory, and a worse-fitting one.

Regional fitting (F2) helps on Census self-response, with a pooled gain over F1 whose interval excludes zero, and does not help on turnout, where the interval spans zero. Both results are reported in full in the companion study. The lesson we draw is not to fit weights to turnout or to Census response. The two proxies reward different weightings, which is what a product-specific, outcome-learned model must accommodate and what no single asserted vector can.

2.3 Ablations: two dimensions do not earn their weight, and the proxies disagree

Table 2 reports leave-one-dimension-out ablations of the full index, with the remaining weights renormalized. Figure S2 (Supporting Information) plots the same values. On Census self-response, removing Accessibility raises ρ from 0.634 to 0.656, and removing Culture raises it to 0.660. Two-fifths

of the index's asserted weight, the whole of Accessibility and the whole education dimension, is on this proxy dead weight or worse. Removing Economics (0.550) or Geography (0.583) damages the score, so the composite is not uniformly hollow. It is specifically mis-weighted. On turnout the picture inverts in a way no single weighting can fix. Removing Culture collapses the score to 0.077, because education carries what little signal the index has. Removing Identity raises it to 0.231, because the age-built dimension works against an age-driven outcome. Removing Geography raises it to 0.218. Paired against the full index on the same resamples, the two Census gains are decisive: removing Accessibility raises ρ by + 0.022 (95% CI [+ 0.011, + 0.033]) and removing Culture by + 0.025 ([+ 0.009, + 0.043]). So is the turnout cost of removing Culture (-0.080, [-0.104, -0.057]). All ten paired contrasts, with Holm adjustment across the ten, are in Table S2 (Supporting Information).

Table 2
Leave-one-dimension-out ablations (Spearman ρ [95% CI], remaining weights renormalized, 3,097 counties per outcome). Intervals are state-clustered bootstrap, 2,000 replications.

Index variant	Census self-response	Turnout
Full GCEIA		
GCEIA (all five dimensions)	0.634 [0.579, 0.680]	0.157 [0.055, 0.248]
One dimension removed		
minus Geography	0.583 [0.513, 0.642]	0.218 [0.116, 0.309]
minus Culture	0.660 [0.605, 0.705]	0.077 [-0.035, 0.181]
minus Economics	0.550 [0.498, 0.595]	0.059 [-0.043, 0.145]
minus Identity	0.629 [0.572, 0.677]	0.231 [0.137, 0.318]
minus Accessibility	0.656 [0.602, 0.701]	0.159 [0.061, 0.247]

The two proxies therefore disagree about the sign of two dimensions, Culture and Geography, and the disagreement is not noise (Table S2). Identity is neutral on the Census proxy, and its removal raises the turnout correlation. A dimension that helps one real response behavior hurts another. This is the empirical core of the paper. For a multi-purpose receptivity index there is no outcome-free resolution of the weights, because receptivity is not one behavior. The weights must be learned per behavior and per product category from data in which the behavior occurred, or the index must be presented, honestly, as a prior.

3. GCEIA-R: a revised, frozen specification

The revision (Table S1, Supporting Information) changes two things, declines two others, and is then frozen.

Geography: density replaces population. Population is an extensive variable. It sums across units, so its percentile measures the size of a county and not the receptivity of its residents. The companion paper showed that this is also the dimension that breaks the index across scales. State and county constructions agree at only ρ = 0.52 on the population dimension, and a fixed fine-scale reference with density in place of population raises that agreement to 0.94 while barely moving the county ranking (Author, 2026a). Density is the intensive counterpart of the same construct, settlement concentration. We validated this change on scale-consistency grounds before the outcome diagnosis existed, and we would make it without the diagnosis. That an index dimension can change meaning with the scale of aggregation is the modifiable areal unit problem (Openshaw, 1984), and indicator-scale interactions of this kind are documented in vulnerability-index construction (Hinojos et al., 2026).

Accessibility: the commute input is dropped. Mean commute time measures labor-market integration, because long commutes mark the counties of a metropolitan labor shed. It does not measure a resident's access to goods, services, or media. It is the wrong construct, and it is also the input whose removal the ablation rewards (Table 2). Accessibility becomes the broadband-subscription percentile alone, the direct access measure the panel contains. The principled second input is drive time to services, which the panel does not contain. It is on the acquisition list (Section 5.5), and until it exists the dimension is one input wide and is labeled as such wherever the index is shown.

Culture is kept, against the ablation. Removing Culture is the single most rewarding deletion on the Census proxy and the single most damaging on turnout (0.077). Education is a receptivity construct with independent standing, and deleting it on the evidence of one proxy would be tuning to that proxy. The same reasoning keeps Identity, whose removal turnout rewards (0.231) and Census leaves essentially unchanged (0.629). Per-outcome learning (Section 5) exists to settle both disputes. An index revised by deleting whatever one proxy dislikes has simply been fitted by hand, badly, to that proxy.

Weights are kept, and renamed. W is unchanged, but its status changes. It is a prior, the starting point from which outcome data may move, and the index's documentation and outputs must say so. We compute percentiles once against the fixed 3,106-county national reference, report scores unrounded, and display tiers with uncertainty bands rather than exact ranks. The free baselines (equal weights, income, population, bachelor's rate) are computed alongside the index permanently, as the yardstick it must beat.

Figure 1 maps the revised index, and Figure S1 (Supporting Information) maps where the revision moves counties. The movement is concentrated where the theory says it should be. Dense metropolitan counties gain on Geography, and counties whose Accessibility rested on short commutes rather than connectivity lose.

4. Evaluation record on the spent proxies

4.1 Correlation and rank record

Table 3 and Fig. 2 place GCEIA-R on the same two proxies, scored by the same protocol. On Census self-response the revision reaches ρ = 0.666 (95% CI [0.616, 0.706]) against the original's 0.634. The paired difference is + 0.032, and its state-clustered bootstrap interval [0.018, + 0.047] excludes zero. The revision also edges past both single-deletion variants of Table 2. On turnout it reaches 0.170 against 0.157 (paired interval spanning zero), remaining far below median income (0.299) and the bachelor's rate (0.346) alone. Equal weights score 0.626 and 0.089. The revision widens the gap over equal weights on the Census proxy (0.666 against 0.626), and on turnout the two intervals overlap.

Table 3
Static indices and fitted models against the two proxies (Spearman ρ [95% CI]). Fitted models are out-of-sample under state-blocked ten-fold cross-validation. Intervals are state-clustered bootstrap, 2,000 replications, 3,097 counties.

Candidate	Census self-response	Turnout
Static		
GCEIA (asserted weights)	0.634 [0.579, 0.680]	0.157 [0.055, 0.248]
GCEIA-R (revised, frozen)	0.666 [0.616, 0.706]	0.170 [0.076, 0.256]
Equal weights	0.626 [0.578, 0.667]	0.089 [-0.012, 0.179]
Median income alone	0.567 [0.500, 0.630]	0.299 [0.201, 0.397]
Bachelor's rate alone	0.429 [0.347, 0.499]	0.346 [0.267, 0.414]
Fitted to the outcome (for scale)		
F1 global linear, five dimensions	0.685 [0.632, 0.737]	0.564 [0.501, 0.628]
F2 regional, shrunk to global	0.737 [0.689, 0.777]	0.609 [0.543, 0.671]
F3 gradient boosting, eight inputs	0.758 [0.715, 0.797]	0.770 [0.728, 0.802]

Table 4 states the same disagreement in the units in which an index is actually consumed, which are ranks. Against the out-of-fold F2 ranking fitted to each proxy, the asserted GCEIA-R ranking correlates at ρ = 0.875 on Census self-response but at only ρ = 0.247 on turnout. The top deciles overlap by 210 of 310 counties on the first proxy and by 30 of 310 on the second. The median county moves 244 ranks under the Census-fitted weights and 791 under the turnout-fitted ones, and on turnout 2,102 of 3,097 counties move 500 ranks or more. Learned weights do not perturb the asserted ranking. On one of the two behaviors they replace it, which is why the revision enters Section 5 as a prior and not as a settled ordering.

Table 4
Rank agreement between the asserted GCEIA-R ranking and the F2 fitted ranking (out-of-fold predictions under the protocol of Section 2.1, with 3,097 counties per proxy and rank 1 as highest).

Statistic	Census self-response	Turnout
Spearman correlation of ranks	0.875	0.247
Kendall's τ-b	0.693	0.162
Top-decile overlap	210 of 310 counties (67.7%)	30 of 310 counties (9.7%)
Counties moving 500 or more ranks	689	2,102
Median absolute rank move	244	791

We state the reading once, plainly, because it constrains everything after it. These proxies informed the diagnosis of Section 2, so Section 4 cannot certify the revision. The Census improvement is consistent with the revision being better, and it is equally consistent with the revision having absorbed, through the diagnosis, information about the Census outcome. We specified the changes on construct and scale-consistency grounds (Section 3), not by searching specifications against these outcomes, and we will evaluate no further change on them. The only clean resolution of the ambiguity is new outcome data scored under rules fixed in advance. That is Section 5. Turnout's verdict needs no such caution. On a participation outcome dominated by age, this family of composites, revised or not, is beaten by two single variables. Any product claim for the index must exclude that behavior or earn its way back on data of its own.

4.2 Pricing the disagreement in households

Rank churn is not outcome loss. Table 4 shows that the asserted and the fitted rankings disagree, but a planner pays in households reached, not in ranks moved. We therefore turn each ranking into a contact plan. The budget is one tenth of the households in the 3,097 matched counties, 12.8 million households, taking ACS population over average household size. Counties are contacted in descending order of the ranking until the budget is spent. The yield is the share of contacted households that self-responded to the 2020 Census, taken as the county self-response rate. The same plan is run on turnout, with residents in place of households and votes cast per resident as the yield. Fitted rankings are the out-of-fold predictions of Section 2.1, so no county is scored by a model that saw its state. Intervals resample states, as in Table 3. We fixed the one-tenth budget before running the plan, to match the top-decile convention of Section 5, and Table S3 reports four budgets so the choice cannot carry the result.

Table 5 gives the result. The asserted index finds a better slice of the country than an untargeted plan does. On Census self-response its yield is 73.9% against 67.7%, which is about 62,000 more responding households per million contacted, or 0.80 million over the whole budget. A plan that contacts the largest counties first does worse than no targeting at all (67.0%). Equal weights give up 1.9 points to the asserted weights (95% CI -3.3 to -0.6), about 240,000 households at this budget, so the asserted weights earn their keep over the neutral choice. What the data do not show is a cost of the asserted weights

large enough to price. The fitted regional model gains 1.7 points over the asserted index, about 220,000 households, and the interval (-0.8 to + 3.8 points, or -110,000 to + 480,000 households) includes zero. GCEIA-R is indistinguishable from GCEIA (-0.5 points, CI -2.5 to + 1.0), so the Census gain of Table 3 does not reappear once the ranking is spent on a concentrated plan. Turnout gives the same pattern at lower stakes. The asserted index reaches 52.6% against 47.7%, the revision and the fitted model sit within a point or two of it, and only the population plan is clearly worse.

Table 5
Contact plans at a budget of one tenth of households (Census) or residents (turnout), in descending order of each ranking. Yield is the responding share of those contacted. Differences are in percentage points against GCEIA, with state-clustered bootstrap intervals (2,000 replications).

Contact plan	Census: yield (%)	Census: difference from GCEIA, points [95% CI]	Turnout: yield (%)	Turnout: difference from GCEIA, points [95% CI]
Untargeted (national average)	67.7	n/a	47.7	n/a
Population size	67.0	-6.9 [-10.7, -4.8]	43.6	-8.9 [-14.8, -5.8]
Equal weights	72.0	-1.9 [-3.3, -0.6]	51.6	-0.9 [-1.9, + 0.5]
GCEIA (asserted weights)	73.9	reference	52.6	reference
GCEIA-R (revised, frozen)	73.4	-0.5 [-2.5, + 1.0]	51.5	-1.1 [-2.1, + 1.2]
Fitted regional model (F2, out-of-fold)	75.7	+ 1.7 [-0.8, + 3.8]	53.1	+ 0.6 [-1.7, + 4.4]

Table S3 repeats the exercise at budgets of 5, 25 and 50 percent. The fitted model clears zero once, on Census self-response at the smallest budget (+ 2.4 points, CI + 0.2 to + 4.7), and its edge is gone by a quarter of households. The pattern is stable. Two readings follow. First, rank disagreement overstates the cost of an asserted weight vector at the top of a ranking. The turnout top deciles overlap by only 30 of 310 counties, yet the yield gap between the asserted and fitted plans is 0.6 points. Whatever the fitted model knows that the asserted one does not, a concentrated plan mostly does not need it, plausibly because the top of an affluence-weighted ranking is high-response almost whichever way the weights fall. Second, this prices the proxies and is not a campaign result. These outcomes informed the diagnosis, the households are estimates, and the yield is a response rate, not a purchase. In reporting the decisive test of Section 5 we will express top-decile capture in households and yield in the same way, as a reporting convention. It changes no gate and no bar of the frozen design.

5. The decisive test, pre-specified

This section is a pre-specification in manuscript form. We wrote it, and froze the accompanying specification, before any campaign outcome data existed. The specification was registered on OSF

Registries on 28 October 2026, before any campaign outcome data were collected. When the test is run, we will report its results against this text whether they pass or not. Publishing the design here is deliberate. It enters the literature as a frozen, citable commitment, fixed in print before the field test it governs has run. We re-scoped the specification once, before any outcome data existed, when a design-based power analysis showed that the original version could not settle the question. What follows is the single specification of record, and the re-scope is documented in the companion power analysis provided with the Supporting Information. Figure 3 summarizes the design.

5.1 Design

The unit is the county, and the design runs in annual cycles. In each cycle we draw 1,280 counties from the 3,106-county frame, stratified by Census division crossed with urbanicity (density tercile). Within strata, counties are randomized to exposure or holdout, so about 640 counties are newly scored per cycle, and a county drawn in several cycles pools its exposures across all of them. Five cycles are scheduled, and cycles repeat beyond the fifth if the gate below is not yet met. The gate, not the calendar, triggers the analysis. We use one offer and one creative set in every exposed county, so that place is the only thing that varies. If creative varied with place, the model would learn copy quality and call it geography. Outcomes are logged per county (impressions, spend, clicks, signups, sales, revenue). Signup or sale is the primary endpoint, and clicks are a weak first signal. A supportive analysis on the click endpoint, whose event volume is roughly five times the primary endpoint's, is pre-specified alongside and cannot substitute for the primary bar.

We attempt no superiority analysis until the pooled data hold at least 2,100 scored areas, at least 15,000 success events (at least seven per scored area), and at least 40 state clusters. Below that gate the data are reported descriptively only. At the declared planning rates the schedule clears the gate in the fifth cycle with about 2,130 scored counties and 15,000 pooled success events. Every candidate ranking (GCEIA-R, the frozen baselines, and the learned model of Section 5.3) is scored with the protocol of Section 2.1.

A design-based power analysis anchored to this program's frozen evaluation record sets the schedule and the gate (companion power analysis). It extends the framework that showed the design's first specification, a single cycle gated at 300 areas and 100 success events, had power below 5% at its then-registered effect. Two measured anchors do the work. Top-decile capture, the primary estimand of Section 5.4, carries its own state-clustering anchor (ICC 0.013 by state-cluster bootstrap on the frozen proxies, an order of magnitude below the correlation statistic's measured 0.09). Its registered floor of 0.125 is the capture-scale expression of the correlation floor of 0.20 (population value 0.137 under the anchored model). At the registered accrual, power is about 0.82 to certify the capture floor and 0.95 to certify a correlation advantage of 0.20, at a joint false-pass rate below 1%. A capture floor of 0.10 is not reachable at 80% power at any accrual inside the county frame (at most 0.76, at ten years and 30,000 events). The floors registered here are the smallest this test can honestly promise.

5.2 Logging schema

We log one append-only row per county-level exposure event. Each row carries an event identifier and timestamp, a hashed account identifier, the industry category, and the county (and ZIP or tract where targeted). It records the index version and the dimension scores, composite, and tier actually shown, the action taken, the creative and offer identifiers, spend, the holdout flag, randomization stratum, and seed, and the outcome counts. It also records the competitive context: rival advertisers or outlets active in the category and county, the auction or list clearing price, and the account's own realized cost per outcome. Creative and offer identifiers are mandatory fields, not metadata. The competitive-context fields are what the decision layer of Section 5.6 is estimated from, and they cannot be reconstructed after the fact. Pooling across accounts happens only at county-by-category aggregate level and only under privacy terms that permit it. That is a launch blocker for the test, not a footnote to it.

5.3 Learning the weights: hierarchical Bayes from the prior

For category c and county i in region r(i), we model the outcome rate with a linear predictor on the five z-scored dimensions, $\eta_i = \Sigma_d w_{c,r,d} \cdot z_{d,i}$, with weights on the simplex through a softmax of latent scores θ. Partial pooling (Gelman & Hill, 2007) supplies what per-user samples cannot. Region scores are drawn around their category, $\theta_{c,r} \sim \text{Normal}(\theta_c, \sigma_{region})$. Category scores are drawn around a global center, $\theta_c \sim \text{Normal}(\theta_0, \sigma_{category})$. And θ_0 is centered on the GCEIA-R prior weights. A category with no outcome data is therefore GCEIA-R exactly, and data move it only as far as the evidence supports. User-level scores form a further level only when a user's own history clears the Section 5.1 gate at user scale. This design retires asserted per-category weight tables. Any weight a product shows must trace to this model or to the prior. We issue forecasts as expected response with intervals, and a point forecast without an interval does not ship. Budget allocation explores by Thompson sampling (Russo et al., 2018) restricted to the prior's top tier, because a small budget cannot fund textbook exploration of places the index already ranks poorly.

5.4 The hero bar

The pass bar, which we call the hero bar, defines what the learned model must do to earn the claim the asserted index could not. On held-out campaign outcomes scored by state-blocked cross-validation, it must win where a ranking is used, at the top, and it must hold the correlation comparison and its calibration. The primary estimand is top-decile capture. With counties ranked by their pooled observed success rate, a ranking's capture is the share of its own top decile that also sits in the outcome's top decile (chance alone captures a tenth). The learned model's capture must exceed every frozen baseline's, with paired intervals excluding zero after Holm correction across the set. The baselines are static GCEIA-R, equal weights, median income rank, population rank, bachelor's-rate rank, and a principal-component index on the raw inputs. The design is powered at a registered capture floor of $\Delta C = 0.125$ (Section 5.1). As the secondary estimand, the paired Spearman advantage under the protocol of Section 2.1 must beat the same baselines on the same terms, with the design powered at a registered floor of $\Delta\rho = 0.20$. And the model's 80% prediction intervals must cover 75–85% of held-out outcomes. All three components are required. Calibration is part of the bar, not a diagnostic, because a ranking that wins on correlation and misstates its uncertainty has not earned a forecaster's claim. Beating proprietary commercial

indices is not the target, since their inputs are not public and the comparison could not be audited. The target is beating the free baselines, on the user's own outcomes, in the open.

5.5 Inputs and resolution

Four input additions are queued in order. The first is age bands and household composition as separate learnable inputs (Section 2.3 shows what a single median hides). The second is drive-time access to retail and services, the principled Accessibility input. The third is spending patterns, with the Bureau of Labor Statistics Consumer Expenditure Survey as the free start. The fourth is competitor and store density from points of interest, since receptivity is relative to what is already present. Finer geography follows the same discipline. We publish tract-level scores under the fixed-reference density specification (national tract distribution, frozen per release) only for categories whose pooled county-level model has passed the Section 5.4 bar, and only with uncertainty bands. Inputs alone never justify finer resolution. Evidence does.

5.6 The decision layer: from scores to strategy

Receptivity is a gross quantity, and decisions are net. A county's value to a firm depends on who else is acting there. Rivals divide the response the index predicts, and their bidding sets the price of reaching it. The final stage of the design is therefore game-theoretic. The players are the firm, choosing entry, bids, and budgets county by county (or staying out), rival firms in the category, which we observe and do not control, and the platform auction that prices attention locally. The payoff is expected net value: learned receptivity multiplied by expected share given the rivals present, minus acquisition cost at the observed clearing price. Expected share takes a discrete-choice (logit) form over the firms present, with learned receptivity as the systematic component. A firm's own outlets enter as rivals to one another, so cannibalization is priced by the same machinery. The output is not a score but a call, enter, avoid, or scale, issued with an interval. A companion study of ZIP-level demand modeling (Author, 2026b) shows why the call and the ranking can part company. In that study, a decision layer of this kind was run on simulated surfaces under stated cost assumptions and returned avoid for every shortlisted ZIP Code Tabulation Area, including those the demand ranking placed first. The result illustrates the rule and is not a finding about any real market, but it shows that a ranking says where response is while a priced decision says what that response is worth. The layer is estimable only from the logged outcomes of Section 5.1 together with the competitive-context fields of Section 5.2, and only for categories whose learned model has passed the Section 5.4 bar. Strategy derived from an unearned ranking is assertion in a second costume. Until the bar is passed the layer is specified and not estimated, and no output of it may be presented as anything else.

For policy and practice, this is the point at which an index stops being a description and becomes an allocation rule. Governments, agencies, firms, and platforms already use composite scores to decide where outreach is sent, where services are sited, and where investment is concentrated. Those decisions are competitive and budget-bound whether the score acknowledges it or not. A receptivity score that ignores rivals and prices systematically overstates the value of the places everyone else can also see.

The decision layer is therefore not an ornament on the index. It is the form in which the index, if its decisive test is passed, would actually be used in practice.

6. Limitations

The decisive limitation is stated in the title of Section 5: the test has not been run. At the time of writing there is no campaign outcome data for this index at all, and nothing in Sections 2 to 4 should be read as evidence about marketing response. The proxies are civic behaviors (voting, answering the Census) whose drivers overlap with consumer receptivity but are not the same. The evaluation of GCEIA-R reuses the diagnostic proxies and is labeled accordingly (Section 4). Accessibility in GCEIA-R is a single-input dimension pending drive-time data. The household counts of Section 4.2 are estimates (ACS population over average household size), and the yields are Census and voting response rates, which bound what a concentrated contact plan can show on these behaviors and not what it will show on purchases. The fitted comparisons of Section 2.2 fit the proxies themselves, including their leaks (especially age into turnout). They bound what outcome-fitting can extract from these inputs on these behaviors, not what it will extract from campaign data. Finally, the hierarchical model's pooling assumes consent and privacy terms that permit county-by-category aggregation across accounts. Where they do not, learning is confined to the global and category levels identifiable from public and first-party aggregate data.

7. Conclusion

Held to outcomes, an asserted receptivity index ties free baselines, loses dimensions it cannot spare, keeps dimensions it cannot justify, and is beaten on both available proxies by models allowed to learn. The correct response is not a better assertion. It is the sequence we followed here: diagnose in the open, revise only what theory indicts, freeze the revision, and pre-specify the test that can promote the index from a prior to a finding or retire the claim. We offer GCEIA-R in that spirit, as the prior with its yardstick attached. For the practitioner the deliverable is narrower and more useful than a claim of superiority. It is a receptivity score whose defects are documented rather than rumored, a revision frozen before its decisive data exist, a disagreement priced in households instead of ranks (smaller at the top of a ranking than the rank churn suggests, and not distinguishable from zero for the revision), and a decision layer that stays switched off until the test is passed, because an allocation rule is only as trustworthy as the validation behind the ranking it reads.

Declarations

Statements and Declarations

Use of artificial intelligence. Claude helped with language and grammar.

Data Availability Statement

All inputs are public: American Community Survey 5-year estimates (U.S. Census Bureau, 2023), the Census Bureau's 2020 self-response rates map data file (U.S. Census Bureau, 2021), and county-level 2020 presidential election results compiled in a public repository (tonmcg, n.d.). County geometry is the Census Bureau's 2023 cartographic boundary file. The trimmed input tables, the outcome files, and the Python code that regenerates every number, table, and figure in this paper from a fixed seed (code/build_gceia_r.py, code/study2_outcome_test.py, code/households.py, code/make_figures.py, code/make_manuscript.py) are provided as Supporting Information and will be deposited in a public repository with a persistent identifier on acceptance. The frozen specification of GCEIA-R (SPEC.md) and the pre-specification text of Section 5 (HERO_TEST_DESIGN.md) accompany the Supporting Information with their freeze date. The specification is registered on OSF Registries (28 October 2026).

Competing Interests

The author develops commercial market-analysis software that implements a version of the GCEIA index evaluated in this paper.

Author Contribution

Nicholas Sokol conceptualized the paper, collected the data, designed the study, selected the methods, implemented and ran the tests, and wrote the paper.

Data Availability

All inputs are public: American Community Survey 5-year estimates (U.S. Census Bureau, 2023), the Census Bureau's 2020 self-response rates map data file (U.S. Census Bureau, 2021), and county-level 2020 presidential election results compiled in a public repository (tonmcg, n.d.). County geometry is the Census Bureau's 2023 cartographic boundary file. The trimmed input tables, the outcome files, and the Python code that regenerates every number, table, and figure in this paper from a fixed seed are provided as Supporting Information and will be deposited in a public repository with a persistent identifier on acceptance. The frozen specification of GCEIA-R and the pre-specification text of Section 5 accompany the Supporting Information with their freeze date. The specification is registered on OSF Registries (28 October 2026), DOI 10.17605/OSF.IO/B49KZ.

References

1. Author. (2026a). [Reference withheld for double-anonymous review]. Unpublished manuscript.
2. Author. (2026b). [Reference withheld for double-anonymous review]. Unpublished manuscript.
3. Bottoms, B., Arbit, J., Lewis, E., et al. (2024). Towards urban place-based resilience modeling: Mixed methods for a flood resilience assessment index. Environment and Planning B: Urban Analytics and

City Science, 51(7), 1563–1580. https://doi.org/10.1177/23998083241243104

4. Cao, Q., Tourre, V., Leduc, T., et al. (2026). Integrating functional accessibility and pedestrian environmental comfort: A spatially explicit framework for urban walkability evaluation. Environment and Planning B: Urban Analytics and City Science. Epub ahead of print 7 January 2026. https://doi.org/10.1177/23998083251413000
5. Feldmeyer, D., Folch, D. C., & Tate, E. (2026). Quantifying the impact of sampling error in composite index construction. Social Indicators Research, 182(1). https://doi.org/10.1007/s11205-026-03827-7
6. Gelman, A., & Hill, J. (2007). Data analysis using regression and multilevel/hierarchical models. Cambridge University Press.
7. Greco, S., Ishizaka, A., Tasiou, M., et al. (2019). On the methodological framework of composite indices: A review of the issues of weighting, aggregation and robustness. Social Indicators Research, 141(1), 61–94. https://doi.org/10.1007/s11205-017-1832-9
8. Hinojos, S., & Grady, C. (2026). An internal validation assessment of scale across composite index model structures. Risk Analysis, 46, e70258. https://doi.org/10.1111/risa.70258
9. Hinojos, S., Roscoe, K., & Grady, C. (2026). Examining the role of indicators and scale in social vulnerability index construction: A comparative geospatial analysis of inductive and hierarchical models. Applied Geography, 186, 103822. https://doi.org/10.1016/j.apgeog.2025.103822
10. Libório, M. P., de Abreu, J. F., Ekel, P. I., et al. (2023). Effect of sub-indicator weighting schemes on the spatial dependence of multidimensional phenomena. Journal of Geographical Systems, 25(2), 185–211. https://doi.org/10.1007/s10109-022-00401-w
11. Magnuszewski, J., Beecham, R., & Burns, L. (2025). The Auckland Urban Liveability Index: A mechanism for quantifying and evaluating modern urban densification. Applied Spatial Analysis and Policy, 18(1). https://doi.org/10.1007/s12061-025-09643-9
12. Norman, P., Lloyd, C., McLennan, D., Ferguson, S., & Catney, G. (2024). 50-year deprivation trajectories: Local area change in England and Wales, 1971–2021. Applied Spatial Analysis and Policy, 17(3), 1183–1208. https://doi.org/10.1007/s12061-024-09583-w
13. Openshaw, S. (1984). The modifiable areal unit problem (Concepts and Techniques in Modern Geography No. 38). Geo Books.
14. Paruolo, P., Saisana, M., & Saltelli, A. (2013). Ratings and rankings: Voodoo or science? Journal of the Royal Statistical Society: Series A, 176(3), 609–634. https://doi.org/10.1111/j.1467-985X.2012.01059.x
15. Punzo, G., Trunfio, M., Castellano, R., et al. (2022). A multi-modelling approach for assessing sustainable tourism. Social Indicators Research, 163(3), 1399–1443. https://doi.org/10.1007/s11205-022-02943-4
16. Russo, D. J., Van Roy, B., Kazerouni, A., et al. (2018). A tutorial on Thompson sampling. Foundations and Trends in Machine Learning, 11(1), 1–96.

17. Saisana, M., Saltelli, A., & Tarantola, S. (2005). Uncertainty and sensitivity analysis techniques as tools for the quality assessment of composite indicators. Journal of the Royal Statistical Society: Series A, 168(2), 307–323. https://doi.org/10.1111/j.1467-985X.2005.00350.x
18. Schmidtlein, M. C., Deutsch, R. C., Piegorsch, W. W., et al. (2008). A sensitivity analysis of the social vulnerability index. Risk Analysis, 28(4), 1099–1114. https://doi.org/10.1111/j.1539-6924.2008.01072.x
19. tonmcg. (n.d.). US county level election results 08–24 [Data set]. GitHub. https://github.com/tonmcg/US_County_Level_Election_Results_08-24
20. U.S. Census Bureau. (2021). 2020 census self-response rates map data file [Data set]. U.S. Department of Commerce.
21. U.S. Census Bureau. (2023). American Community Survey 5-year estimates [Data set]. U.S. Department of Commerce.

Figures

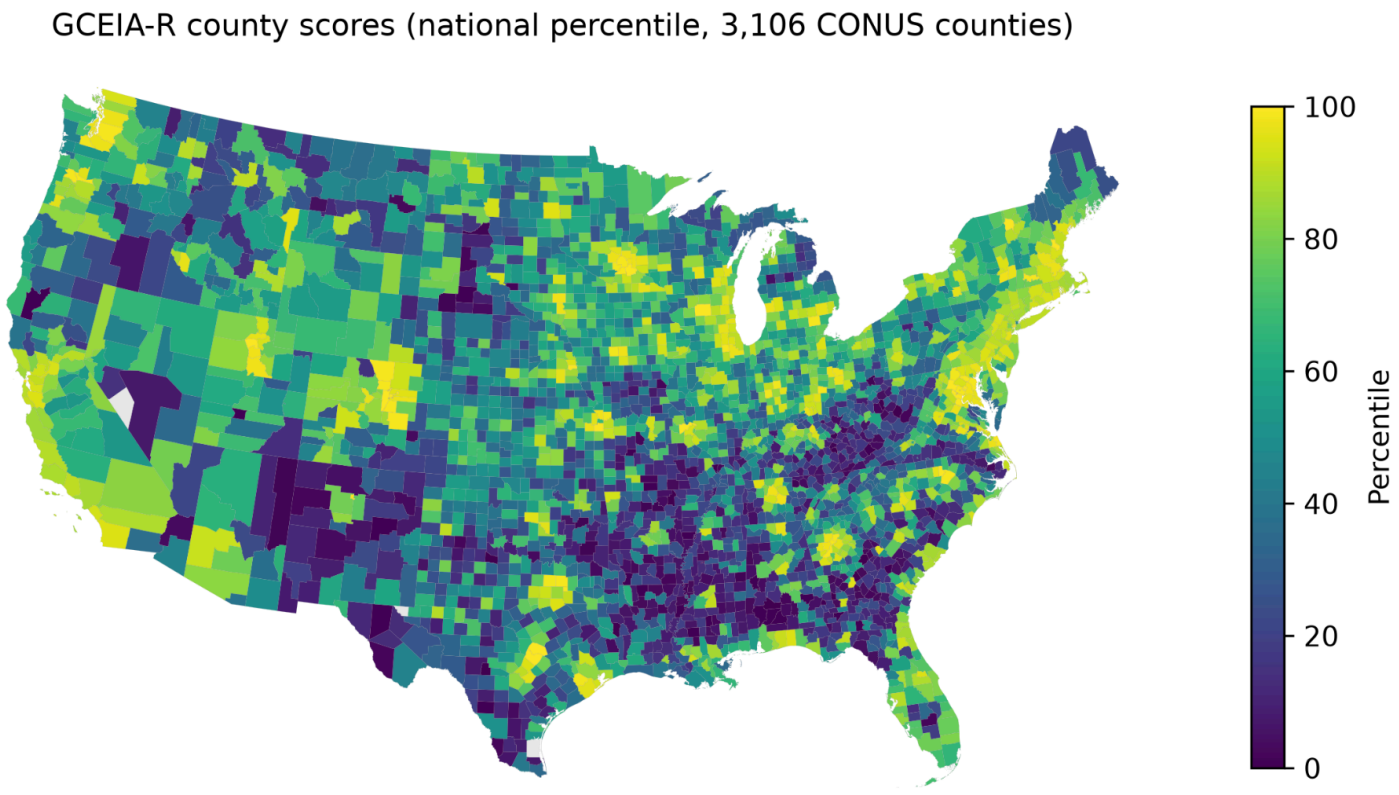


Figure 1

GCEIA-R county scores (national percentile).

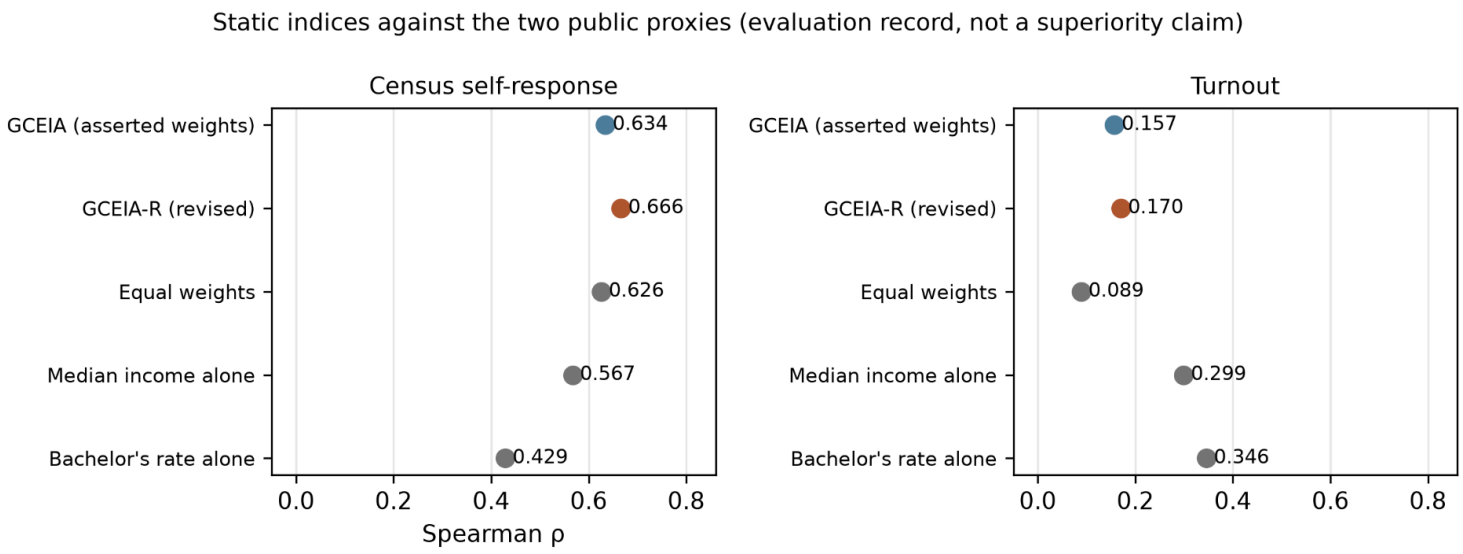


Figure 2

Static indices against the two proxies: the evaluation record of Table 3.

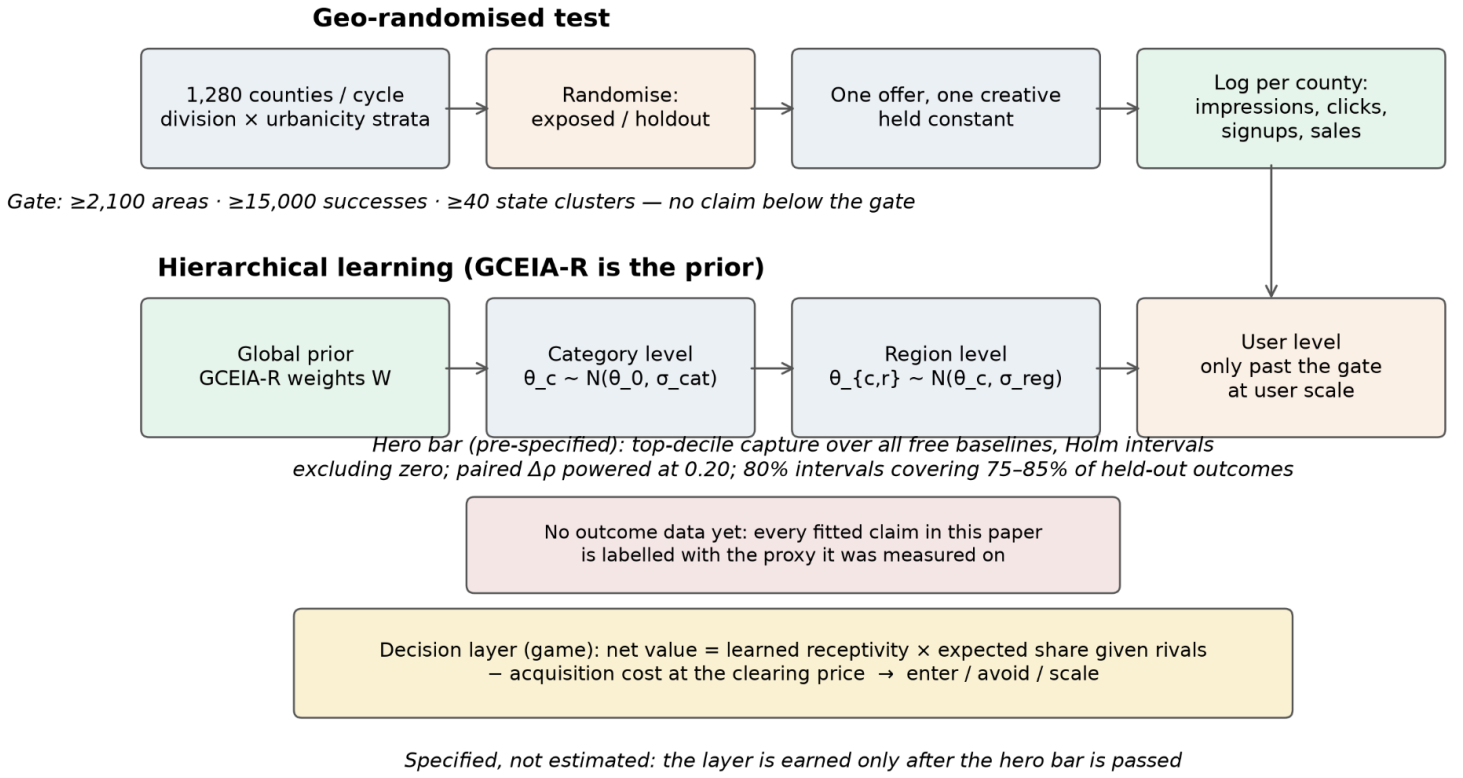


Figure 3

The pre-specified design: multi-year geo-randomized accrual, gate, hierarchical learning, hero bar.