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.

Calibration and Configurational Recovery in NCA and fsQCA: A Monte Carlo Benchmark for Public-Service Research in Vietnam

Thanh Vu Nguyen
Thai Nguyen University of Economics and Business Administration
Tuan Anh Tran
tuananh@tueba.edu.vn

Thai Nguyen University of Economics and Business Administration
Thi Thu Hoai Pham
Thuongmai University

Research Article

Keywords: Necessary Condition Analysis, fuzzy-set Qualitative Comparative Analysis, Monte Carlo simulation, calibration sensitivity, scope definition, public-service research

Posted Date: November 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11118533/v1

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

Additional Declarations: No competing interests reported.

Abstract

Calibration and scope choices determine what set-theoretic estimands represent, yet applied studies often treat them as neutral preprocessing. We benchmark the recovery and false-detection behaviour of Necessary Condition Analysis (NCA) and fuzzy-set Qualitative Comparative Analysis (fsQCA) using a public-service measurement framework motivated by five rural services in Vietnam. The evaluation uses 456,000 synthetic datasets from 456 scenarios with 1,000 replications each. Scenarios vary sample size, measurement error, condition correlation, cellwise missingness, outcome resolution, missingness mechanism and two data-generating processes: a two-condition gate and a substitutive non-gate. Across the 432-scenario missing-completely-at-random factorial, fsQCA calibrated with substantive 2/3/4 anchors recovered the programmed gate in 0.749 of replications and falsely detected the pair in 0.015 of non-gate replications. Raw-score NCA yielded rates of 0.673 and 0.087, respectively. NCA applied to 2/3/4 fuzzy memberships was conservative (0.049 recovery), whereas percentile-calibrated fsQCA was less specific (0.519 recovery; 0.162 false detection). In matched scenarios, increasing sample size improved semantic-anchor fsQCA recovery but reduced raw NCA recovery as observations filled the ceiling region. These results show that calibration anchors and NCA scope definitions are part of the estimand. Applied studies should preregister external anchors, report scope and algorithmic non-results, and test recovery against a plausible counterfactual data-generating process. The benchmark evaluates analytical workflows; it does not estimate how Vietnamese residents assess public services.

1. Introduction

Citizen satisfaction is consequential for the legitimacy and adaptive capacity of public services, but it is not a generic attitude detached from what a service produces. Expectations, experienced performance, performance signals and service design jointly shape evaluation (Bellé et al. 2025; Chan et al. 2021; Favero and Kim 2021; Wang and Kim 2024; Zhang et al. 2022). Frontline encounters matter especially in person-intensive services (Petrovsky et al. 2023), whereas continuity and technical integrity are more immediate constraints for water, sanitation and transport. Vietnam's programme spans these domains (Prime Minister of Vietnam 2022; World Bank 2022), and recent evidence treats satisfaction as service- and programme-specific (Le et al. 2026; Pham et al. 2023; Tao et al. 2025; Tung et al. 2022).

These features make set-theoretic methods attractive. Necessary Condition Analysis (NCA) asks whether a minimum level of a condition constrains the maximum attainable outcome (Dul 2016, 2024a). Fuzzy-set Qualitative Comparative Analysis (fsQCA) asks whether combinations of conditions are sufficient or necessary, and it accommodates conjunction, equifinality and asymmetry (Oana et al. 2021; Ragin 2008; Schneider and Wagemann 2012). Used together, the two methods appear to offer a sequence from bottlenecks to alternative pathways (Vis and Dul 2018). Agreement between them, however, is not confirmation: their estimands, geometry and sensitivity to calibration differ. Recent methodological work has therefore emphasised explicit design logic, calibration, algorithm choice and robustness rather than mechanical solution reporting (Baumgartner 2022; Baumgartner and Thiem 2020; Dul et al. 2023; Duşa 2022; Greckhamer et al. 2018; Oana and Schneider 2024; Pappas and Woodside 2021; Parkkinen and Baumgartner 2023).

The stakes of those choices are not confined to methodology. Studies of public services use NCA to argue that a capability sets a minimum requirement for satisfactory provision, and use fsQCA to argue that a particular combination of capabilities is sufficient; such claims inform decisions about staffing, maintenance and investment. Yet applied studies rarely test whether a claim survives a defensible change in calibration anchors or in the scope rectangle used to define empty ceiling space, even though both choices determine what the resulting quantity means. When one set of responses can be read as evidence of a necessary capability floor, or as no evidence at all, depending only on those choices, the problem is interpretive before it is statistical. In programme settings it also becomes budgetary. Vietnam's New Rural Development Programme is a case in point: capability and satisfaction measures collected at commune level feed into local planning, and national citizen-experience indices are increasingly used to compare provinces (CECODES et al. 2025; Le et al. 2026; Prime Minister of Vietnam 2022). An unexamined calibration or scope can therefore change which service a local government is told to prioritise. This article treats calibration and scope as objects of study rather than as preprocessing, and asks how far conclusions travel when those choices change.

The motivating dataset for this study was a transparent, single-seed synthetic benchmark modelled on five rural public-service domains. Its generator directly included a two-condition minimum operator. Recovering that gate was useful for checking code, but it could not test the theory that motivated the generator. The original workflow also applied NCA to fuzzy memberships, treated statistical significance as sufficient even for very small effects, and assessed only one sample, so it could not address power, false detection, measurement error or robustness. It illustrates a wider problem: a clean result in one designed dataset may reflect the design more than the method's capacity to distinguish a gate from a plausible non-gate.

This article therefore asks three methodological questions. First, how do NCA and fsQCA recover a programmed two-condition gate across alternative calibrations, scopes and sample sizes? Second, how often do they report the same pair when the central conditions are substitutive rather than conjunctural? Third, how do measurement error, condition correlation, missingness and outcome resolution change these rates?

The study contributes a transparent benchmark that separates recovery under a programmed gate from false detection under a substitutive process. It embeds that benchmark in an ADEMP design, evaluates calibration and scope as components of the estimand, and translates the results into a preregistration protocol for applied public-service research. The service-capability architecture supplies a disciplined measurement context, but it remains a prospective framework. No result is interpreted as evidence about Vietnamese residents.

2. Conceptual basis and evaluation targets

2.1 Satisfaction as a service-specific capability judgement

Public Service Logic locates value in use and within service ecosystems rather than in an isolated transaction (Engen et al. 2021; Osborne 2020; Osborne et al. 2022). Citizens, however, are not simply

private customers: collective obligations, equal access and institutional responsibility remain central (Sønderskov and Rønning 2021), so satisfaction should be read alongside objective safety, equity and statutory performance rather than as a substitute for them (Barbera et al. 2025).

We organise perceived provision into four capability families. Access and continuity (AC) captures availability, reliability and regularity. Structural and technical adequacy (ST) captures facilities, equipment and technical condition. Human capability (HC) captures professional competence and problem solving. Relational responsiveness (RR) captures empathy, communication and response to users. The families synthesise rather than replicate conventional service-quality dimensions (Parasuraman et al. 1988); they also let the same vocabulary travel across domains without assuming equal causal roles.

The proposed boundary is the service production system. Health and education are person-intensive: HC and RR are plausible joint constraints because expertise must be communicated and enacted through an encounter. Water, sanitation and rural transport are infrastructure-network services: AC and ST are plausible joint constraints because access without technical adequacy, or technical adequacy without continuity, cannot deliver a dependable service. Table 1 maps the five domains, and Fig. 1 separates this architecture from the benchmark that evaluates analytical recovery.

Table 1
Service-specific capability mapping and programmed central pair

Service	System	AC	ST	HC	RR	Pair
Health	Person	Reliability; response	Facilities	Competence	Empathy	HC–RR
Education	Person	Access	Facilities	Teachers	Environment and response	HC–RR
Water	Network	Reliability	Facilities	Service competence	Response	AC–ST
Sanitation	Network	Frequency; operations	Equipment	Workforce	Response	AC–ST
Transport	Network	Reliability	Technical quality	Service competence	Response	AC–ST

Note: AC = access and continuity; ST = structural and technical adequacy; HC = human capability; RR = relational responsiveness. The pair is installed by the synthetic data-generating process and is not an empirical finding.

The mapping is a set of prospective propositions for field research, not a result of the synthetic analysis. The benchmark deliberately assigns HC-RR as the central pair for health and education and AC-ST for water, sanitation and transport. Its substantive value is to define what future data should measure; its methodological value is that the true programmed structure is known.

Note

AC = access and continuity; ST = structural and technical adequacy; HC = human capability; RR = relational responsiveness. Solid arrows show the complementary analytical roles of NCA and fsQCA. The dashed arrow indicates that the service-production system conditions which capability pair is theoretically central. The framework is prospective and does not represent estimated empirical effects in Vietnam.

2.2 Why NCA and fsQCA may diverge

NCA operationalises necessity through empty space above a ceiling line, and its effect size depends on the observed ceiling and on the rectangle used to define the scope (Dul et al. 2020, 2023). Raw continuous scores are sufficient for NCA; fuzzy calibration is not a requirement. Applying NCA to calibrated memberships is nevertheless a useful sensitivity exercise when the transformation and scope are explicit: capping memberships at 0.047 and 0.953 while imposing a theoretical [0,1] scope mechanically leaves an empty top strip, whereas the empirical scope removes that strip but may leave few distinct outcome levels. Effect size, permutation evidence and a pre-specified practical threshold must therefore be read together (Dul 2024a).

fsQCA evaluates set-subset relations and minimises sufficient truth-table rows, so its results depend on concept definition, calibration, frequency and consistency thresholds, limited diversity and the minimisation algorithm (Rutten 2022; Swiatczak 2022; Thiem and Mkrtchyan 2024; Thomann and Maggetti 2020). Large-N applications still require substantive knowledge, and empirical regularity alone does not establish causal power (Finn 2022; Rutten 2023). A solution may be stable under one perturbation and unstable under another. Fit statistics and case classification should therefore accompany formula comparison (Oana and Schneider 2024). Calibration is particularly consequential: anchors tied to response meaning preserve an external set definition, whereas percentile anchors define membership relative to each sample. The latter are included here as a stress test, not as a recommended default.

Four evaluation targets follow. Under the gate data-generating process (DGP), a useful workflow should recover both central conditions; under the substitutive DGP, reporting the same pair is a false pair detection. A discriminating method should combine high gate recovery with low non-gate detection, and conclusions should not reverse merely because of a defensible alternative anchor, modest measurement noise or realistic incomplete data. These are performance targets, not tests of the capability propositions.

3. Methods

3.1 Design and disclosure

The study has two stages. Stage 1 re-audits the original dataset of 560 synthetic household-style profiles generated with seed 20260919. Place labels and identifiers in that file are fictional structural devices; no respondent was surveyed and no administrative record was used. Stage 1 checks whether conclusions change when NCA is run on raw scores, fuzzy memberships, empirical and theoretical scopes, and when practical relevance is required. Stage 2 is a multi-scenario Monte Carlo study reported using the aims, data-generating mechanism, estimands, methods and performance-measures (ADEMP) framework (Morris et

al. 2019; Siepe et al. 2025); it evaluates repeated-sampling recovery and false detection. Because the DGP contains the target structure, recovery is evidence about the analytical pipeline only and cannot corroborate the public-service theory. The full ADEMP specification, decision rules and scenario grid are in Online Resource 1 (Table S1).

3.2 Seeded illustration and programmed structure

The illustration contains four observed family scores and a separate five-category satisfaction outcome for each of five services. Let C1 and C2 denote the assigned central pair, P1 and P2 the remaining pair, and all family scores be rescaled to [0,1]. Latent satisfaction was generated as

$$\eta \;=\; 0.12 \;+\; 0.58\,\min(C_1,\,C_2) \;+\; 0.20\,\max(P_1,\,P_2) \;+\; 0.10\,\mathrm{mean}(AC,\,ST,\,HC,\,RR) \;+\; \epsilon$$

The minimum operator installs the capability gate, and the maximum operator permits substitution between peripheral conditions. Consequently, observing C1 and C2 in NCA or fsQCA solutions is recoverability, not an independent hypothesis test. The synthetic file's internal-consistency coefficients show only that its item generator behaved as designed; they are not questionnaire validation.

3.3 Monte Carlo data-generating mechanisms

For each replication, four latent capabilities were drawn from an equicorrelated multivariate normal distribution with unit variances, common mean 0.35 and pairwise correlation ρ, then transformed by the standard normal cumulative distribution function. Four ordinal items per capability were generated by adding normally distributed measurement error, mapping to the 1–5 scale and averaging. The gate DGP used the equation above with ε ~ N(0, 0.055²). The non-gate DGP replaced min(C1, C2) with max(C1, C2):

$$\eta \;=\; 0.12 \;+\; 0.58\,\max(C_1,\,C_2) \;+\; 0.20\,\max(P_1,\,P_2) \;+\; 0.10\,\mathrm{mean}(AC,\,ST,\,HC,\,RR) \;+\; \epsilon$$

Both DGPs therefore retain positive average associations and peripheral equifinality, but only the first contains a joint central gate. Latent satisfaction was clipped to [0,1]. The observed outcome was either one rounded five-category item or the mean of three independently perturbed five-category items.

The full factorial crossed four sample sizes, three measurement-error levels, three condition correlations, three cellwise missing rates, two outcome measures and two DGPs, yielding 432 missing-completely-at-random (MCAR) scenarios. Cellwise missingness was applied independently to the four conditions and the outcome, so expected complete-case retention at m = 0.20 was $(1 - m)^5 = 0.328$ rather than 0.80. Twenty-four targeted missing-at-random (MAR) stress scenarios held measurement error at 0.15 and ρ at 0.30 and raised missingness for profiles with lower latent satisfaction. Table 2 summarises the 456 scenarios; each used 1,000 replications, producing 456,000 datasets. Replication seeds were derived from master seed 20260920, the scenario identifier and the replication number. Analyses used complete cases and returned an analysis failure when fewer than 30 remained.

Table 2
Monte Carlo design (456 scenarios; 1,000 replications per scenario)

Factor	Levels	Implementation
Sample size	100, 250, 560, 1,000	Full MCAR factorial; MAR: 100, 560, 1,000
Measurement error SD	0.05, 0.15, 0.30	MAR fixed at 0.15
Condition correlation	0.00, 0.30, 0.60	Equicorrelation; MAR fixed at 0.30
Cellwise missing rate	0, 0.10, 0.20	Complete-case analysis
Outcome	Single 5-point; mean of 3 items	Ordinal rounding after perturbation
DGP	Gate; substitution	$\min(C_1,C_2)$ versus $\max(C_1,C_2)$
Mechanism	432 MCAR + 24 MAR scenarios	MAR weighted toward low latent satisfaction
Replications	1,000 per scenario	456,000 total; master seed 20260920

Note: The full MCAR factorial is 4 × 3 × 3 × 3 × 2 × 2 = 432 scenarios. The targeted MAR block adds 24 scenarios.

3.4 Calibration, estimation and performance measures

Direct calibration used a piecewise logit transformation with slope 3 and three alternatives: substantive 2/3/4 anchors, full-range 1/3/5 anchors and sample percentiles P10/P50/P90. Values outside the lower and upper anchors were capped, yielding memberships of 0.047, 0.500 and 0.953 at the three anchor points; ties were moved to adjacent observed categories where necessary.

Ceiling envelopment-free disposal hull (CE-FDH) NCA was estimated in five ways: raw observed scores with empirical scope; 2/3/4 fuzzy memberships with empirical scope; 1/3/5 fuzzy memberships with empirical scope; percentile memberships with empirical scope; and 2/3/4 memberships with the theoretical [0,1] × [0,1] scope. A condition was detected only when $d \geq 0.10$ and a one-sided randomisation $p < 0.05$. This rule prevents a very small but statistically detectable empty area from being labelled substantively necessary in a large sample. The simulation used 99 random permutations after effect-size screening, and the seeded audit used 9,999. Pair recovery required both central conditions to meet the rule; any detected peripheral condition was recorded separately.

fsQCA used the three calibrated datasets. Cases at exactly 0.5 were assigned to set presence when constructing truth-table rows. A sufficient row required consistency ≥ 0.80, proportional reduction in inconsistency (PRI) ≥ 0.65 and frequency ≥ 3 for fewer than 200 complete cases, ≥ 5 for 200–799, and ≥ 10 for 800 or more. Exact Boolean minimisation produced complex solutions without using logical remainders, and recovery was deliberately strict: both central conditions had to appear positively in every term of every equally optimal solution. A model with no qualifying row or no solution was an algorithmic non-result.

The primary gate-recovery and false pair-detection rates are unconditional: an algorithmic non-result counts as a failure to recover and as no positive detection. Output rates and conditional rates among returned results are reported separately in the supplement, which avoids silently deleting difficult replications. Wilson 95% intervals accompany scenario rates; with 1,000 replications, the largest nominal Monte Carlo standard error for an unconditional proportion is 0.0158.

The implementation included an exact four-condition Boolean minimiser and a CE-FDH routine. Automated checks confirmed 456 unique scenarios, 1,000 unique replications per scenario and 456,000 unique keys, and the 2/3/4 theoretical-scope NCA effects reproduced the archived seeded results within 0.000477. All code, seed rules, the scenario grid and replication-level metrics are in the anonymised reproducibility package.

4. Results

4.1 Re-audit of the single-seed illustration

The re-audit overturns the strongest reading of the original NCA results (Online Resource 1, Table S2). On raw scores with empirical scope, all five intended pairs were recovered, but seven of ten peripheral conditions were also detected, so the apparent sensitivity came with poor specificity. With fuzzy 2/3/4 memberships and empirical scope, only three of ten central conditions met $d \geq 0.10$ and no service pair was recovered. With the theoretical [0,1] scope, five central conditions and only the education pair met the combined criterion. The archived significance-only statement that all ten central tests were necessary was therefore driven by p-values, even though five effects were below 0.10; the common peripheral effect of about 0.045 under theoretical scope reflects the capped membership range rather than substantive necessity.

The full-range 1/3/5 calibration recovered all five pairs but also detected three peripheral conditions under empirical scope and four under theoretical scope, whereas percentile calibration recovered three pairs without peripheral detections. No single variant simultaneously justified a general necessity claim and cleanly separated all intended from peripheral conditions. Stage 1 should thus be read as a pipeline audit whose answer depends on scale and scope, not as support for the programmed propositions.

4.2 Pooled discrimination across the MCAR factorial

Across the 432-scenario MCAR factorial, semantic-anchor fsQCA produced the best overall discrimination (Table 3; Fig. 2). The 2/3/4 specification recovered the programmed pair in 0.749 of gate replications and falsely reported it in 0.015 of substitutive replications; 1/3/5 produced 0.724 and 0.013. Raw-score NCA recovered 0.673 but its false pair-detection rate was 0.087. Its balanced accuracy was 0.793, below 0.867 for fsQCA with 2/3/4 anchors.

Calibration did not move both methods in the same direction. NCA on 2/3/4 memberships with empirical scope was highly conservative: gate recovery was 0.049 and false detection 0.000. The theoretical scope increased recovery only to 0.082, and the 1/3/5 NCA variant reached 0.505 recovery with 0.018 false

detection. In contrast, percentile fsQCA recovered 0.519 but falsely reported the pair in 0.162 of non-gate datasets. Pooled rates are marginal averages over deliberately heterogeneous scenarios and should be used to compare profiles, not as universal performance constants.

Table 3
Unconditional performance across the 432-scenario MCAR factorial

Method	Gate recovery [95% CI]	False pair [95% CI]	Output G/N	Balanced acc.
NCA raw / empirical	0.673 [0.671, 0.675]	0.087 [0.086, 0.088]	0.979/0.979	0.793
NCA 2/3/4 / empirical	0.049 [0.048, 0.050]	0.000 [0.000, 0.000]	0.979/0.979	0.525
NCA 1/3/5 / empirical	0.505 [0.503, 0.507]	0.018 [0.018, 0.019]	0.979/0.979	0.743
NCA percentile / empirical	0.149 [0.147, 0.150]	0.014 [0.013, 0.014]	0.979/0.978	0.568
NCA 2/3/4 / theoretical	0.082 [0.081, 0.083]	0.000 [0.000, 0.000]	0.979/0.979	0.541
fsQCA 2/3/4	0.749 [0.747, 0.750]	0.015 [0.014, 0.015]	0.976/0.979	0.867
fsQCA 1/3/5	0.724 [0.722, 0.725]	0.013 [0.013, 0.013]	0.979/0.979	0.855
fsQCA percentile	0.519 [0.517, 0.521]	0.162 [0.161, 0.164]	0.964/0.953	0.678

Note: Each DGP contributes 216,000 replications. Non-results count as non-recovery or no positive detection. Output G/N gives the proportions returning a usable gate/non-gate result. Balanced accuracy = [recovery + (1 − false detection)]/2.

Note

Rates are marginal across sample size, measurement error, correlation, missingness and outcome resolution. The shaded area marks false detection ≤ 0.05. E = empirical NCA scope; T = theoretical NCA scope.

4.3 Sample size exposes different failure mechanisms

Figure 3 holds measurement error at 0.15, ρ at 0.30, missingness at zero and uses the single-item outcome. Raw NCA gate recovery fell from 0.853 at n = 100 to 0.720 at n = 250, 0.294 at n = 560 and 0.082 at n = 1,000, while its false pair detection fell from 0.032 to zero. Small samples more often left apparently empty ceiling regions, and additional observations filled those regions even under the true but noisy gate.

By contrast, fsQCA with 2/3/4 anchors rose from 0.568 to 0.847, 0.984 and 0.999, with zero false pair detection in these four matched cells. Percentile fsQCA also recovered more often as n grew (0.271 to

0.768), but its non-gate false pair rate rose from 0.081 to 0.313, so more data strengthened both the desired and the undesired percentile-defined pattern. Fuzzy 2/3/4 NCA with empirical scope recovered no pair in the matched cells, because the discretised, capped memberships rarely produced two effects at or above 0.10.

Note

Measurement error = 0.15; ρ = 0.30; no missingness; single five-category outcome; 1,000 replications per DGP and sample size. Error bars show Wilson 95% confidence intervals.

4.4 Measurement, dependence and missingness

Increasing measurement error from 0.05 to 0.30 reduced gate recovery for raw NCA from 0.974 to 0.265 and for fsQCA with 2/3/4 anchors from 0.885 to 0.637. The corresponding semantic fsQCA false rate remained between 0.012 and 0.018, whereas percentile fsQCA false detection increased from 0.062 to 0.288. For raw NCA, false detection decreased with noise because ceiling effects weakened under both DGPs; lower false detection therefore did not indicate better discrimination. Condition correlation had a different effect. As ρ increased from 0 to 0.60, raw NCA recovery rose from 0.614 to 0.732, but non-gate false pair detection rose from 0.005 to 0.210. Semantic fsQCA recovery fell from 0.820 to 0.652, while its false detection rose more modestly from 0.003 to 0.031. Correlation can thus make separate necessary-condition tests appear jointly persuasive even when the DGP permits substitution.

At 20% cellwise MCAR, some n = 100 scenarios retained too few complete cases, and the maximum scenario-level analysis-failure rate was 0.279. Pooled output rates remained between 0.953 and 0.980, but semantic fsQCA recovery declined from 0.822 with no missingness to 0.665 at m = 0.20. The three-item outcome improved raw NCA recovery (0.808 versus 0.538) while also increasing its non-gate detection (0.114 versus 0.060); semantic fsQCA was comparatively insensitive to outcome resolution (0.743 versus 0.754). In the targeted MAR comparison, semantic fsQCA recovery was 0.658 versus 0.767 in matched MCAR cells, whereas percentile fsQCA moved in the opposite direction (0.758 versus 0.584), showing that distribution-dependent anchors can interact with selective missingness. Complete results, conditional rates and uncertainty intervals are in Online Resource 1 (Tables S3–S6).

5. Discussion

5.1 What the benchmark does and does not establish

The benchmark separates recovery of a programmed structure from evidence for a substantive theory. Because the original DGP installs a gate through min(C1, C2), recovering C1 and C2 only shows that the method can identify the encoded structure under the simulated conditions. The single-seed illustration remains useful for detecting implementation and reporting problems, but method-performance claims require repeated samples and comparison with the substitutive DGP. The Monte Carlo conclusions therefore apply only to the simulated design space.

The results also reject a simple claim that one method is generally more robust. Semantic-anchor fsQCA best discriminated the two DGPs in this study, but its recovery deteriorated with noise, dependence and incomplete cases. Raw NCA was more sensitive in some small or low-noise cells, yet it frequently detected peripheral or substitutive conditions. Fuzzy NCA could be extremely conservative, and percentile fsQCA appeared increasingly powerful with sample size while becoming less specific. NCA and fsQCA should therefore not be presented as validating each other: a necessary-condition ceiling and a sufficient configurational subset are different empirical objects (Dul et al. 2023; Vis and Dul 2018).

5.2 Why calibration and scope change the answer

Three mechanisms explain the divergence. First, NCA depends on empty ceiling space. With stochastic measurement, larger samples are more likely to include a case that violates a formerly empty region, so raw CE-FDH recovery can decline even when a noisy gate generated the data. This is consistent with the importance of sampling design for NCA (Dul 2024b). Second, fuzzy calibration changes both geometry and resolution. The 2/3/4 transformation collapses a five-category outcome to three membership values after capping; under empirical scope this leaves little ceiling area, whereas under theoretical scope the unobserved strip above 0.953 contributes mechanically to d. Reporting the scope and the ceiling plot is therefore essential rather than optional.

Third, percentile calibration imposes relative distinctions. In fsQCA it can manufacture apparently coherent sufficient rows even when the central conditions substitute, particularly as n stabilises the sample quantiles. Percentiles are not a straw-man primary calibration here: they are a deliberately adverse, commonly encountered sensitivity strategy, and adding the substantively plausible 1/3/5 alternative shows that sensitivity is not unique to them. A credible application should choose anchors from concept meaning before inspecting solutions, and then vary only alternatives that preserve that meaning (Oana et al. 2021; Oana and Schneider 2024).

5.3 Implications for public-service research and practice

For applied public-service research, the results imply a concrete reporting standard. The same simulated responses supported opposite necessity conclusions when anchors or NCA scope changed, and a substitutive process produced apparently coherent pairs under distribution-based calibration. Studies should therefore report calibration anchors, the NCA scope rectangle, ceiling plots, algorithmic non-results and at least one substantively defensible sensitivity analysis. If the interpretation remains unchanged, the claim is stronger; if it changes, the analysis should present the alternative conclusions rather than select one post hoc.

The capability architecture remains a proposal awaiting field evidence. Its benefit is disciplined comparability: all services are described through AC, ST, HC and RR, while the central pair changes with the production system. This avoids both extremes of treating every service as incomparable and collapsing all services into a single satisfaction score. It aligns with evidence that core, facilitating and supporting service attributes interact (Chan et al. 2021), that frontline conditions shape citizen encounters (Petrovsky et al. 2023), and that expectations update with information and performance (van der Voet and van den Bekerom 2025).

For Vietnam's New Rural Development Programme, the framework could guide a preregistered multi-service survey, but the present coefficients and solutions must not guide resource allocation. A field study should combine perceived capabilities with objective indicators such as service availability, water safety, collection frequency, road condition and staffing. Recent programme research demonstrates the importance of participation, local context and interdependent determinants (Le et al. 2026; Tung et al. 2022), and the Provincial Governance and Public Administration Performance Index illustrates the feasibility and value of disaggregated citizen-experience measures (CECODES et al. 2025). Because such instruments increasingly inform provincial comparison and local planning, calibration choices left unreported become, in effect, unreported policy choices.

Regression or structural equation modelling would be complementary, not a direct performance control. Those methods estimate average net associations and measurement relations, whereas NCA and fsQCA target ceiling constraints and set-subset configurations. Because both simulated DGPs produce positive average associations, a conventional additive regression can fit associations without distinguishing necessity from substitution. A field study should therefore use confirmatory factor analysis or structural equation modelling to validate measurement, regression for average marginal relations, and set-theoretic methods only for preregistered necessity and configuration questions; disagreement must be interpreted against the estimand rather than counted as failed confirmation (Richter et al. 2020).

5.4 Limitations and field-validation protocol

The simulation is broader than the original benchmark but intentionally finite. It uses four equicorrelated capabilities, two functional forms, fixed weights, normally distributed errors, ordinal rounding and complete-case analysis. Other gates, interactions, floor effects, clustered samples, differential item functioning and non-ignorable missingness may behave differently. The custom implementation was checked against archived outputs and fully disclosed, but independent replication with the NCA, QCA or SetMethods software ecosystems would further strengthen portability. Ninety-nine permutations make the simulation feasible and resolve the 0.05 decision, but give coarser p-values than the 9,999 used for the seeded audit. Finally, the false-pair criterion is tailored to the programmed pair and is not a universal false-positive measure for every possible formula.

A field validation should be preregistered before outcome inspection. It should specify service eligibility and recall periods; justify external anchors; run NCA first on raw construct scores with a declared scope; specify fsQCA directional expectations, thresholds and treatment of remainders; and replicate across provinces or waves. The full checklist is provided in Online Resource 1 (Section S5), and data, code and an immutable analysis plan should be deposited in a persistent repository subject to ethical and legal constraints, so that theoretical risk is separated from analytical flexibility.

6. Conclusion

Across 456,000 simulated datasets, semantic-anchor fsQCA most consistently combined recovery and specificity, raw NCA showed a sample-size-dependent ceiling trade-off, fuzzy NCA depended strongly on calibration and scope, and percentile fsQCA produced substantial non-gate detections. These results do

not show which capabilities determine satisfaction in Vietnam. They show that necessity and configurational claims depend on explicit scale definitions, NCA scope, a plausible counterfactual DGP and repeated-sampling evidence. Treating those choices as part of the estimand makes set-theoretic public-service research more transparent and easier to audit.

Declarations

Funding.

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing interests. The authors have no relevant financial or non-financial interests to disclose.

Ethics approval. Not applicable. The study uses algorithmically generated data only and involves no human participants, no human data and no administrative records.

Consent to participate. Not applicable.

Consent for publication. Not applicable.

Data availability. The study uses only synthetic data. An anonymised reproducibility package containing the generator, master and replication seed rules, the full scenario grid, replication-level key metrics, aggregate tables, validation checks and figure scripts will be uploaded as a separate supplementary file through the submission system. A permanent repository DOI will be added after the journal's blinded-review requirements are confirmed.

Code availability. All analysis code will be included in the anonymised reproducibility package described above.

Materials availability. Not applicable.

Author contributions. Author contributions are listed on the separate title page to preserve anonymity during review.

Use of generative artificial intelligence. Generative-AI tools supported language editing, code scaffolding and editorial revision. The authors reviewed and verified the analytical code, numerical results, references and final text and take full responsibility for the article.

Author Contribution

N.T.V.: Conceptualization, Methodology, Software, Formal analysis, Writing – original draft, VisualizationT.T.A.: Methodology, Validation, Writing – review & editingP.T.T.H.: Supervision, Project administration, Writing – review & editing

Data Availability

The datasets generated during and/or analysed during the current study are not publicly available due to the double-anonymous peer review model, but the anonymised reproducibility package is available in the supplementary files. This package contains the synthetic generator, master and replication seed rules, the complete scenario grid, replication-level key metrics, aggregate tables, validation checks and figure scripts. The package contains no human-participant data. A permanent repository DOI will be deposited and added upon acceptance of the manuscript.

References

1. Barbera, C., Sicilia, M., Steccolini, I.: Exploring government–citizen interaction in public service performance assessment: trade-offs, synergies, and critical issues. Public Adm. **103**, 814–827 (2025). https://doi.org/10.1111/padm.13068
2. Baumgartner, M.: Qualitative Comparative Analysis and robust sufficiency. Qual. Quant. **56**, 1939–1963 (2022). https://doi.org/10.1007/s11135-021-01157-z
3. Baumgartner, M., Thiem, A.: Often trusted but never (properly) tested: evaluating Qualitative Comparative Analysis. Sociol. Methods Res. **49**, 279–311 (2020). https://doi.org/10.1177/0049124117701487
4. Bellé, N., Belardinelli, P., Cucciniello, M., Nasi, G.: Experimental evidence on the determinants of citizens' expectations toward public services. Public Adm. Rev. **85**, 176–191 (2025). https://doi.org/10.1111/puar.13775
5. Analytics, C.E.C.O.D.E.S.R.-T., United Nations Development Programme: : The 2024 Viet Nam Provincial Governance and Public Administration Performance Index: measuring citizens' experiences. UNDP Viet Nam, Hanoi (2025). https://www.undp.org/vietnam/publications/2024-provincial-governance-and-public-administration-performance-index-papi-report. Accessed 20 September 2026
6. Chan, F.K.Y., Thong, J.Y.L., Brown, S.A., Venkatesh, V.: Service design and citizen satisfaction with e-government services: a multidimensional perspective. Public Adm. Rev. **81**, 874–894 (2021). https://doi.org/10.1111/puar.13308
7. Dul, J.: Necessary Condition Analysis (NCA): logic and methodology of necessary but not sufficient causality. Organizational Res. Methods. **19**, 10–52 (2016). https://doi.org/10.1177/1094428115584005
8. Dul, J.: A different causal perspective with Necessary Condition Analysis. J. Bus. Res. **177**, 114618 (2024a). https://doi.org/10.1016/j.jbusres.2024.114618
9. Dul, J.: How to sample in Necessary Condition Analysis. Eur. J. Int. Manage. (2024b). https://doi.org/10.1504/EJIM.2024.10063098
10. Dul, J., van der Laan, E., Kuik, R.: A statistical significance test for Necessary Condition Analysis. Organizational Res. Methods. **23**, 385–395 (2020). https://doi.org/10.1177/1094428118795272

11. Dul, J., Hauff, S., Bouncken, R.B.: Necessary Condition Analysis (NCA): review of research topics and guidelines for good practice. Rev. Managerial Sci. **17**, 683–714 (2023). https://doi.org/10.1007/s11846-023-00628-x
12. Duşa, A.: Critical tension: sufficiency and parsimony in QCA. Sociol. Methods Res. **51**, 541–565 (2022). https://doi.org/10.1177/0049124119882456
13. Engen, M., Fransson, M., Quist, J., Skålén, P.: Continuing the development of the public service logic: a study of value co-destruction in public services. Public. Manage. Rev. **23**, 886–905 (2021). https://doi.org/10.1080/14719037.2020.1720354
14. Favero, N., Kim, M.: Everything is relative: how citizens form and use expectations in evaluating services. J. Public Adm. Res. Theor. **31**, 561–577 (2021). https://doi.org/10.1093/jopart/muaa048
15. Finn, V.: A qualitative assessment of QCA: method stretching in large-N studies and temporality. Qual. Quant. **56**, 3815–3830 (2022). https://doi.org/10.1007/s11135-021-01278-5
16. Greckhamer, T., Furnari, S., Fiss, P.C., Aguilera, R.V.: Studying configurations with Qualitative Comparative Analysis: best practices in strategy and organization research. Strategic Organ. **16**, 482–495 (2018). https://doi.org/10.1177/1476127018786487
17. Le, H.D., Ly, L.T.K., Hoang, V.M., Tan, D.D., Huong, N.T.M., Huong, M.T.L.: Determinants of success in Vietnam's national target program on new rural development: a Bayesian belief network analysis. Qual. Quant. **60**, 8935–8962 (2026). https://doi.org/10.1007/s11135-026-02618-z
18. Morris, T.P., White, I.R., Crowther, M.J.: Using simulation studies to evaluate statistical methods. Stat. Med. **38**, 2074–2102 (2019). https://doi.org/10.1002/sim.8086
19. Oana, I.-E., Schneider, C.Q.: A robustness test protocol for applied QCA: theory and R software application. Sociol. Methods Res. **53**, 57–88 (2024). https://doi.org/10.1177/00491241211036158
20. Oana, I.-E., Schneider, C.Q., Thomann, E.: Qualitative Comparative Analysis Using R: A Beginner's Guide. Cambridge University Press, Cambridge (2021). https://doi.org/10.1017/9781009006781
21. Osborne, S.P.: Public Service Logic: Creating Value for Public Service Users, Citizens, and Society Through Public Service Delivery. Routledge, Abingdon (2020)
22. Osborne, S.P., Powell, M., Cui, T., Strokosch, K.: Value creation in the public service ecosystem: an integrative framework. Public Adm. Rev. **82**, 634–645 (2022). https://doi.org/10.1111/puar.13474
23. Pappas, I.O., Woodside, A.G.: Fuzzy-set Qualitative Comparative Analysis (fsQCA): guidelines for research practice in information systems and marketing. Int. J. Inf. Manag. **58**, 102310 (2021). https://doi.org/10.1016/j.ijinfomgt.2021.102310
24. Parasuraman, A., Zeithaml, V.A., Berry, L.L.: SERVQUAL: a multiple-item scale for measuring consumer perceptions of service quality. J. Retail. **64**, 12–40 (1988)
25. Parkkinen, V.-P., Baumgartner, M.: Robustness and model selection in configurational causal modeling. Sociol. Methods Res. **52**, 176–208 (2023). https://doi.org/10.1177/0049124120986200
26. Petrovsky, N., Xin, G., Yu, J.: Job satisfaction and citizen satisfaction with street-level bureaucrats: is there a satisfaction mirror? J. Public Adm. Res. Theor. **33**, 279–295 (2023). https://doi.org/10.1093/jopart/muac022

27. Pham, L., Limbu, Y.B., Le, M.T.T., Nguyen, N.L.: E-government service quality, perceived value, satisfaction, and loyalty: evidence from a newly emerging country. J. Public. Policy. **43**, 812–833 (2023). https://doi.org/10.1017/S0143814X23000296
28. Prime Minister of Vietnam: Decision No. 263/QD-TTg approving the National Target Programme on New Rural Development for 2021–2025. (2022). https://vanban.chinhphu.vn/?docid=205378&pageid=27160&tagid=4&type=1 Accessed 20 September 2026
29. Ragin, C.C.: Redesigning Social Inquiry: Fuzzy Sets and Beyond. University of Chicago Press, Chicago, IL (2008)
30. Richter, N.F., Schubring, S., Hauff, S., Ringle, C.M., Sarstedt, M.: When predictors of outcomes are necessary: guidelines for the combined use of PLS-SEM and NCA. Industrial Manage. Data Syst. **120**, 2243–2267 (2020). https://doi.org/10.1108/IMDS-11-2019-0638
31. Rutten, R.: Applying and assessing large-N QCA: causality and robustness from a critical realist perspective. Sociol. Methods Res. **51**, 1211–1243 (2022). https://doi.org/10.1177/0049124120914955
32. Rutten, R.: Uncertainty, possibility, and causal power in QCA. Sociol. Methods Res. **52**, 1707–1736 (2023). https://doi.org/10.1177/00491241211031268
33. Schneider, C.Q., Wagemann, C.: Set-Theoretic Methods for the Social Sciences: A Guide to Qualitative Comparative Analysis. Cambridge University Press, Cambridge (2012)
34. Siepe, B.S., Bartoš, F., Morris, T.P., Boulesteix, A.-L., Heck, D.W., Pawel, S.: Simulation studies for methodological research in psychology: a standardized template for planning, preregistration, and reporting. Psychol. Methods. (2025). https://doi.org/10.1037/met0000695
35. Sønderskov, M., Rønning, R.: Public service logic: an appropriate recipe for improving serviceness in the public sector? Administrative Sci. **11**, 64 (2021). https://doi.org/10.3390/admsci11030064
36. Swiatczak, M.D.: Different algorithms, different models. Qual. Quant. **56**, 1913–1937 (2022). https://doi.org/10.1007/s11135-021-01193-9
37. Tao, X., Makarchev, N., Le, D.A., Yao, B., Chinh, D.T., Xiao, C.: Citizen satisfaction with public services: a multi-province study in Vietnam. Australian J. Public. Adm. **84**, 804–821 (2025). https://doi.org/10.1111/1467-8500.12688
38. Thiem, A., Mkrtchyan, L.: Case-to-factor ratios and model specification in Qualitative Comparative Analysis. Field Methods. **36**, 52–68 (2024). https://doi.org/10.1177/1525822X231159458
39. Thomann, E., Maggetti, M.: Designing research with Qualitative Comparative Analysis (QCA): approaches, challenges, and tools. Sociol. Methods Res. **49**, 356–386 (2020). https://doi.org/10.1177/0049124117729700
40. Tung, D.T., Diem, L.T.T., Luan, D.X., Linh, N.H.K.: The National Target Program for New Rural Development in Vietnam: an understanding of people's participation and its determinants. Sustainability. **14**, 12140 (2022). https://doi.org/10.3390/su141912140
41. van der Voet, J., van den Bekerom, P.: Performance information, expectations and satisfaction with public service delivery in the context of co-production initiatives. Public. Manage. Rev. **27**, 1157–1178 (2025). https://doi.org/10.1080/14719037.2024.2392131

42. Vis, B., Dul, J.: Analyzing relationships of necessity not just in kind but also in degree: complementing fsQCA with NCA. Sociol. Methods Res. **47**, 872–899 (2018). https://doi.org/10.1177/0049124115626179
43. Wang, W., Kim, T.K.: Do government performance signals affect citizen satisfaction? J. Policy Anal. Manag. **43**, 846–870 (2024). https://doi.org/10.1002/pam.22597
44. World Bank: National Targeted Programs for New Rural Development and Sustainable Poverty Reduction Support Program: implementation completion and results report (Report No. ICR00005693). World Bank, Washington, DC (2022)
45. Zhang, J., Chen, W., Petrovsky, N., Walker, R.M.: The expectancy-disconfirmation model and citizen satisfaction with public services: a meta-analysis and an agenda for best practice. Public Adm. Rev. **82**, 147–159 (2022). https://doi.org/10.1111/puar.13368

Figures

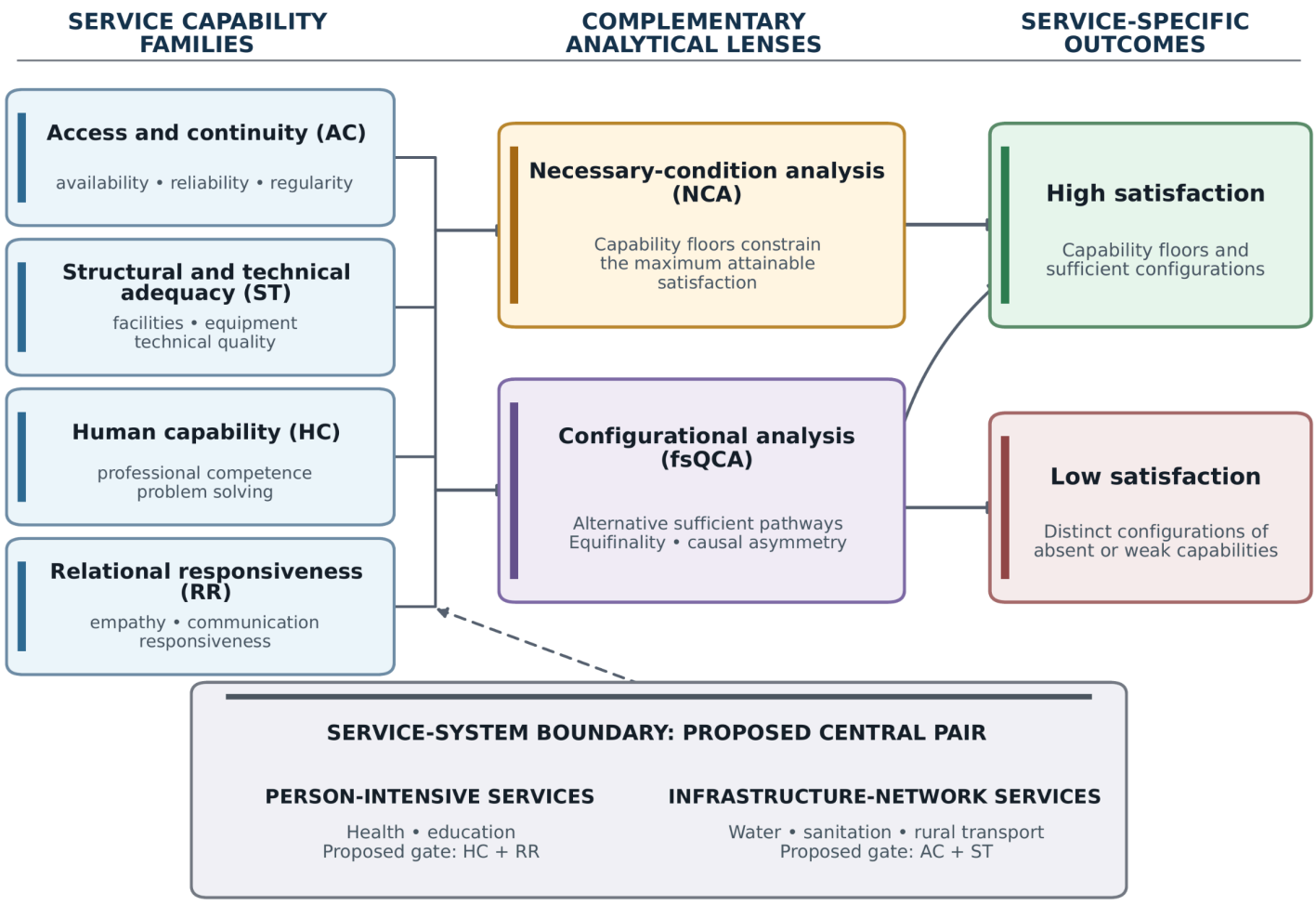


Figure 1

Theory-driven capability architecture and benchmark logic

Note: AC = access and continuity; ST = structural and technical adequacy; HC = human capability; RR = relational responsiveness. Solid arrows show the complementary analytical roles of NCA and fsQCA. The dashed arrow indicates that the service-production system conditions which capability pair is theoretically central. The framework is prospective and does not represent estimated empirical effects in Vietnam.

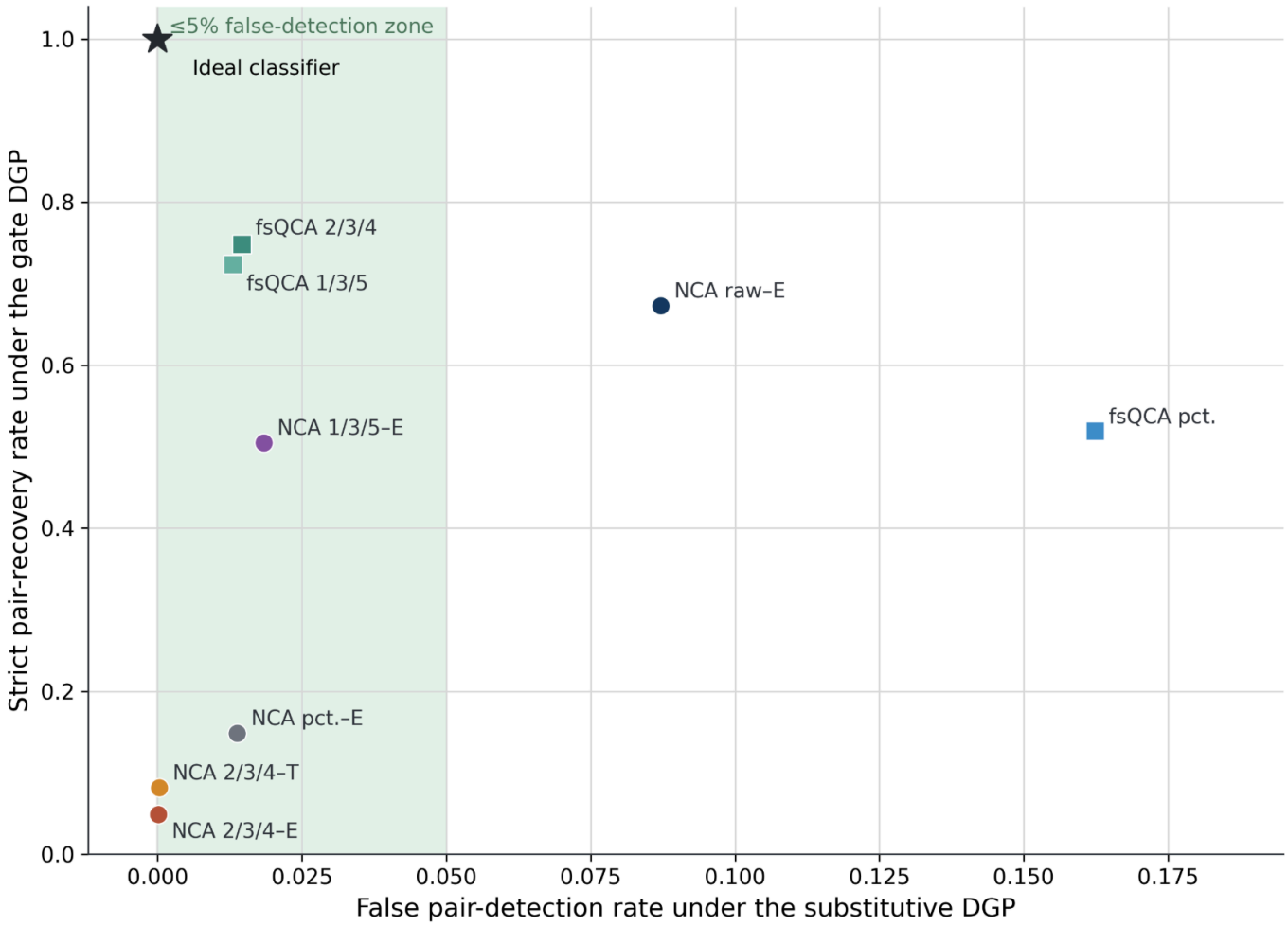


Figure 2

Method discrimination across the 432-scenario MCAR factorial

Note: Rates are marginal across sample size, measurement error, correlation, missingness and outcome resolution. The shaded area marks false detection ≤ 0.05. E = empirical NCA scope; T = theoretical NCA scope.

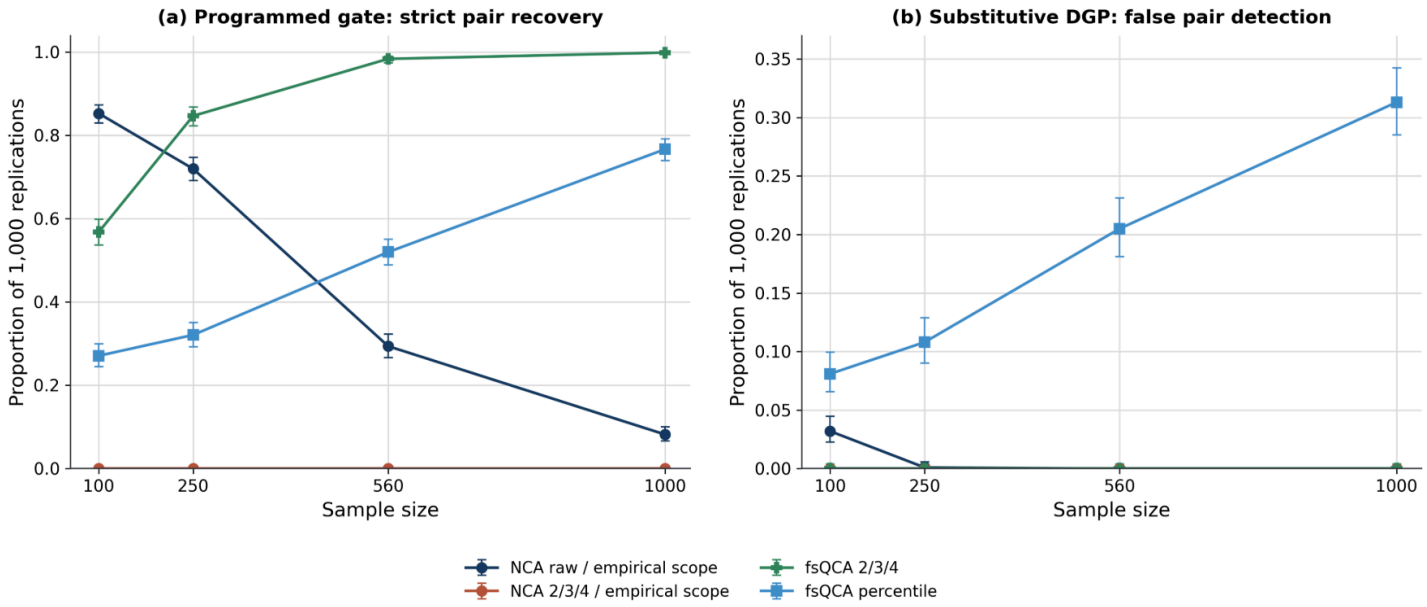


Figure 3

Recovery and false detection in the matched baseline design

Note: Measurement error = 0.15; ρ = 0.30; no missingness; single five-category outcome; 1,000 replications per DGP and sample size. Error bars show Wilson 95% confidence intervals.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- QandQOnlineResource.docx