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.

Extraction-Confidence-Weighted Calibration (ECWC): A Reliability-Weighted Correction for Fuzzy-Set Qualitative Comparative Analysis

Devan Cantrell Addison-Turner

daddisonturner@stanford.edu

Stanford University

Research Article

Keywords: fsQCA, calibration, measurement reliability, configurational methods, robustness testing

Posted Date: November 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11077090/v1

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

Additional Declarations: No competing interests reported.

Extraction-Confidence-Weighted Calibration (ECWC): A Reliability-Weighted Correction for Fuzzy-Set Qualitative Comparative Analysis

Abstract

Fuzzy-set Qualitative Comparative Analysis (fsQCA) calibrates set membership from raw scores, but existing calibration treats all sources as equally reliable; a 34-study audit finds none reports a calibration reliability statistic. I introduce Extraction-Confidence-Weighted Calibration (ECWC), shrinking a case's calibrated membership toward the crossover in proportion to its source's unreliability. The paper's central contribution is a closed-form boundary condition governing when this shrinkage helps versus hurts: simulation establishes that it reduces calibration error only within roughly 8–15 points of the crossover and increases it beyond that range, a pattern the closed-form derivation corroborates qualitatively under a different error criterion. Applied uniformly, the correction is on average worse than no correction; a gated version, restricted to the helps-region, recovers 95.9–97.2% of this cost, and I recommend it over the uniform version wherever the added derivation step is feasible. Four real-data applications across three domains test this pattern, with mixed but broadly consistent results and no ground truth to arbitrate between corrected and gated estimates. ECWC applies wherever fsQCA combines sources of unequal reliability.

Keywords: *fsQCA; calibration; measurement reliability; configurational methods; robustness testing*

1. Introduction

Fuzzy-Set Qualitative Comparative Analysis (fsQCA) identifies which combinations of conditions are necessary or sufficient for an outcome; originating in comparative politics, it has since spread well beyond that origin into individual-level technology-adoption research and across sustainability research as the field embraces configurational explanations over strictly linear ones, with recent applications in construction digitalization (Wei and Liu 2026), green development behavior (Li et al. 2025b), enterprise digital transformation (Zhou et al. 2024), and green technological innovation effectiveness (Liu and Wei 2026), alongside a growing technology-adoption literature (Abbasi et al. 2022) and public-health applications (Wang and Luo 2025). Across both literatures, most applied fsQCA work calibrates every case's raw score into fuzzy-set membership using a single, uniform procedure.

That uniform treatment sits uneasily with how these datasets are built. A single fsQCA analysis might combine life-cycle-assessment figures, sensor data, expert coding, and automated extraction; a technology-adoption study might combine self-reported constructs (Abbasi et al. 2022) with system-logged or automated behavioral-trace data; an umbrella review of 36 systematic reviews finds the same pattern in nature-exposure research, where greenspace studies rely on measures from satellite-derived indices to self-reported contact (Bryer et al. 2025), all sources of unequal reliability standard calibration cannot represent. Recent fsQCA studies report calibration consensus reached only "through discussion," with no inter-rater statistic (Wei and Liu 2026; Shafaghat et al. 2026). I then audited fsQCA studies in two rounds, 34 in total across construction, sustainability, and cross-domain applications (full study list and citations in Online Resource 1,

Supplementary Note S16; **Table 1**'s caption; limitations in Section 4.5). None of the 34 reported a formal inter-rater reliability statistic for calibration, and even the weaker check of anchor-sensitivity was uncommon, occasionally producing a materially different configuration under an alternative anchor set (**Table 1**).

Existing methodological literature diagnoses this problem with precision but has not resolved it: Skaaning (2011), Krogslund et al. (2015), Hug (2013), and Maggetti and Levi-Faur (2013) established that fsQCA is sensitive to calibration and coding decisions, and later sensitivity-diagnostics tools (Oana and Schneider 2024; Toyoda 2026) test how much a solution would change under hypothetical coding error once calibration is already finished, not whether calibration itself should reflect it. The closest prior calibration advances (Legewie 2017; Olsen 2022; Mendel and Korjani 2015, 2018; Batista Hernandez et al. 2026; Dușa 2019) address related but distinct problems (how each falls short is detailed in Online Resource 1, Supplementary Note S17) and none weights by source-level reliability.

No published fsQCA calibration method weights a case's membership by its source-level reliability; I ask whether doing so reduces calibration error, and how far that benefit extends before it becomes a cost. This paper makes three contributions, plus a fourth, more exploratory discussion of scope; of the three, the second is the most consequential: the boundary condition governing when it helps rather than hurts, derived and tested here for the first time. First, I formalize Extraction-Confidence-Weighted Calibration (ECWC) (Section 2.2), a simple, no-new-data adaptation of Kelley's (1927) classical true-score correction from psychometrics (Maassen 2000). Second, I derive the boundary condition governing when that shrinkage helps versus hurts in exact closed form (Section 2.3), the first closed-form location of a reliability-validity tradeoff long recognized in measurement theory only qualitatively (Cronbach and Meehl 1955). Third, I validate ECWC on four real, previously published datasets (Sections 3.2–3.5), none requiring new data, together diagnosing the boundary condition's selectivity rather than uniformly large practical gains. Beyond these three, Section 4.4 discusses ECWC's broader relevance wherever fsQCA combines heterogeneous, unequally reliable data.

2. Materials and Methods

Table 1 completes the audit picture introduced above: of the second, larger round's 28 studies, 2 (7%) reported a genuine calibration-anchor sensitivity check, a further 6 (21%) tested a different, more common parameter instead (the consistency or case-frequency cutoff, not the calibration anchors), and the remaining 20 (71%) reported neither. One partial exception: a single study reported a formal inter-rater reliability statistic for a separate coding task elsewhere in its pipeline, but not for calibrating its fsQCA conditions or outcome, confirming the gap is specific to calibration, not to reliability practice generally.

Table 1. Outcomes for the 28 fsQCA studies whose calibration/methods sections could be fully read, out of 40 candidates identified via public search across 10-plus application domains (12 inaccessible due to publisher paywalls or access barriers). Row 1 is an independent question from rows 2–4, which partition the same 28 studies by calibration-uncertainty/robustness practice (2 + 6 + 20 = 28); one additional study, not counted in row 1, reported an inter-rater statistic for a different data-pipeline stage, not for fsQCA calibration itself. Percentages are each rounded independently and sum to 99%, not 100%.

Outcome (out of 28 studies checked)	Studies	%
Reported a formal inter-rater/inter-coder reliability statistic for fsQCA calibration	0	0%
Reported a calibration-anchor/breakpoint sensitivity check	2	7%
Reported a robustness check of an analytic parameter only (consistency or case-frequency cutoff), not calibration anchors	6	21%
Reported no reliability or sensitivity check of any kind	20	71%

2.1. Standard Calibration

I use Ragin's direct (log-odds) calibration method (Ragin 2008), the field's standard procedure and the baseline ECWC extends. Three qualitative anchors are specified for each raw variable: full non-membership, the crossover point (maximum ambiguity, membership = 0.5), and full membership. A logistic transformation maps each raw score onto the [0,1] membership scale such that a score at the full-membership anchor calibrates to 0.953 and a score at the full-non-membership anchor calibrates to 0.047, the field's conventional near-certainty thresholds.

Two further quantities, defined here for readers outside fsQCA, are used throughout. Necessity consistency measures how consistently a condition is present whenever the outcome is present (0-to-1 scale; conventional necessity threshold above 0.90). Sufficiency consistency measures how consistently the outcome is present whenever a condition or configuration is present (same scale; conventional sufficiency threshold above 0.80). Coverage measures how much of the outcome's occurrence a condition or configuration accounts for, capturing empirical relevance rather than logical strength: a condition can clear a strict consistency threshold while covering only a small share of cases, or vice versa. Because ECWC shrinks a case's calibrated membership before these coefficients are computed, a shift in membership can move a necessity or sufficiency score across a threshold without any change to the underlying data, the specific concern Sections 3.2–3.5 test directly.

2.2. Extraction-Confidence-Weighted Calibration (ECWC)

ECWC adds one step after standard calibration. For a case with standard calibrated membership m and an associated reliability weight $r \in [0,1]$, the ECWC-adjusted membership m' is given by **Equation (1)**:

$$m' = 0.5 + (m - 0.5) \times r \quad \textbf{(1)}$$

A case calibrated from a fully reliable source (r = 1) is left unchanged; one from a source of no reliability (r = 0) is pulled entirely to the crossover, reflecting that its true membership is unknown.

The reliability weight r can be drawn from information a study has already collected: inter-rater or pipeline-versus-human-coder agreement, or a researcher's own calibration-uncertainty estimates (Section 2.6); it extends directly into existing QCA robustness-testing protocols (Oana and Schneider 2024), which can report solution stability with the weighting on and off.

This shrinkage formula is not new: it is structurally identical to Kelley's (1927) classical true-score correction from psychometrics, $T' = \rho X + (1 - \rho)M$, which shrinks an observed test score X toward the group mean M in proportion to its unreliability $(1 - \rho)$, ρ a reliability coefficient on [0,1] (Maassen 2000). Substituting the fsQCA crossover (0.5) for M, and a case- or source-level extraction-confidence weight for ρ, recovers **Equation (1)** exactly. What is new is not the shrinkage principle, nearly a century old, but its adaptation to fuzzy-set calibration: anchoring the target to the QCA crossover rather than a sample mean, sourcing the weight from extraction or coding confidence rather than test-retest reliability, and characterizing, in Section 3.1, precisely when the correction helps and when it hurts.

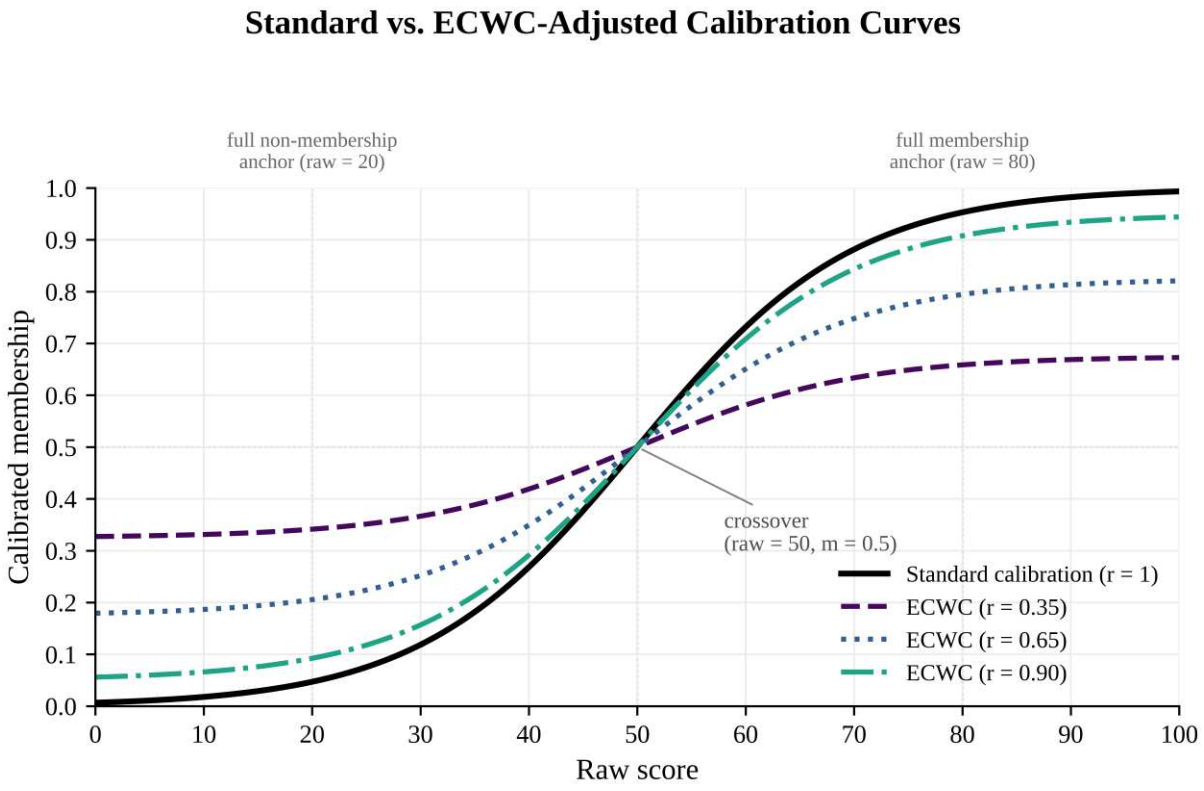


Fig. 1 Standard calibration curve versus ECWC-adjusted calibration curves at three reliability levels (r = 0.35, 0.65, 0.90), as a function of raw score. Lower reliability shrinks attainable membership values toward the crossover (0.5); r = 1 reproduces standard calibration exactly

Figure 1 illustrates this mechanism: the standard calibration curve (r = 1) alongside the ECWC-adjusted curve at three reliability levels. Because **Equation (1)** scales each point's distance from the crossover by r without moving the crossover itself, every curve passes through the same

crossover point (raw = 50, m = 0.5) with the three raw-score anchors fixed on the x-axis. Standard calibration reaches 0.047 and 0.953 at the full non-membership and full-membership anchors (raw = 20, 80), a spread of 0.906; ECWC narrows that spread to exactly r times its standard value: 0.317 at r = 0.35, 0.589 at r = 0.65, 0.815 at r = 0.90: the visible flattening is nothing more than that fixed ratio traced across the raw-score range.

2.3. Simulation: Characterizing the Error-Reduction Boundary

To test whether ECWC reduces calibration error, I simulated a single case with a known, noise-free true raw score T and reliability r, generating many independent noisy "observed" raw scores and comparing the resulting standard and ECWC-adjusted memberships to the true calibrated membership. I swept T across the full raw-score range (10 to 90, steps of 5) at five reliability levels (r = 0.35, 0.50, 0.65, 0.80, 0.90) under Gaussian noise ($\sigma = 30 \times (1 - r)$, clipped to the raw-score range), with 20,000 independent draws per value of T (seed = 42, NumPy default_rng). To check whether the pattern reflects reliability alone rather than noise shape, I repeated the primary r = 0.65 sweep under four further variance-matched noise models (two skew-normal, Azzalini shape parameter = ±4; one bounded uniform; one Student's t, 3 df) and ran the full 5 × 5 grid of all 25 combinations.

For each condition, I computed a 95% Confidence Interval (CI) on the interpolated breakeven distance (the distance from the crossover at which ECWC's benefit turns to a cost) via 200,000 Monte Carlo draws, and derived a closed-form breakeven distance analytically via a second-order (delta-method) linearization of the calibration function, the same family behind Willink's (2026) metrology result, under the Mean-Squared-Error (MSE) criterion his result uses, yielding the defining equation (full derivation in Online Resource 1, Supplementary Note S14):

$$D = (1 + r)(\mu^2 + V) / (2\mu) \quad \textbf{(2)}$$

For this manuscript's logistic calibration this reduces to a cubic in D^2, solved exactly by the classical trigonometric method for a depressed cubic (see Data availability), matching a numerical root-finder to machine precision, while the delta-method approximation itself tracks exact numerical integration (Gauss–Hermite quadrature) to within 3.0–3.3% (Section 3.1). **Equation (2)**'s D converts, via the calibration function's inverse, to a raw-score-equivalent distance, compared against the Monte Carlo breakeven distances in Section 3.1 (**Table 2** visualizes r = 0.65; **Table 3** the full sweep); the two need not coincide exactly, since they use different error criteria (Section 4.1, 4.5). The derivation generalizes beyond the logistic form, validated for probit and complementary log-log links against exact adaptive-quadrature integration under both criteria (Online Resource 1). I used **Table 3**'s own Monte Carlo breakeven distance, not this analytical form, to gate the original single-condition simulation, applying ECWC only within that reliability level's breakeven distance of the crossover, otherwise leaving the case at standard calibration (**Table 4**).

Finally, I tested whether correlated measurement error across jointly-calibrated conditions changes this picture: a three-condition conjunction (fuzzy-AND) simulation, sharing a true raw score and reliability level across conditions at the same reliability levels, T-grid, seed, and noise scale as the primary sweep (20,000 replications per reliability/correlation combination), with per-condition noise drawn from a shared latent factor plus an independent residual whose correlation (κ) took three values (0, 0.5, 1); at κ = 1 every condition receives an identical draw, so the result reduces exactly to the single-condition case, validating against **Table 4**. This check uses only standard and uniform ECWC, not adaptive gating.

The four datasets in Sections 2.4–2.7 were chosen to test **Equation (1)** against the different forms a reliability weight actually takes in practice, not for topical variety: the Introduction's audit (**Table 1**) found fsQCA studies almost never report a formal reliability statistic for calibration, so validating ECWC against only one kind of signal would leave open whether a usable weight exists at all. Sections 2.4 and 2.5 use composite reliability and Cronbach's alpha, already a routine byproduct of each source study's own hybrid Partial Least Squares Structural Equation Modeling (PLS-SEM)/fsQCA design. Section 2.6 uses reliability elicited directly from the original data's author, for a dataset whose publication reported none. Section 2.7 constructs a weight from the data's own sampling structure when neither other source is available, the situation the audit suggests most fsQCA researchers actually face.

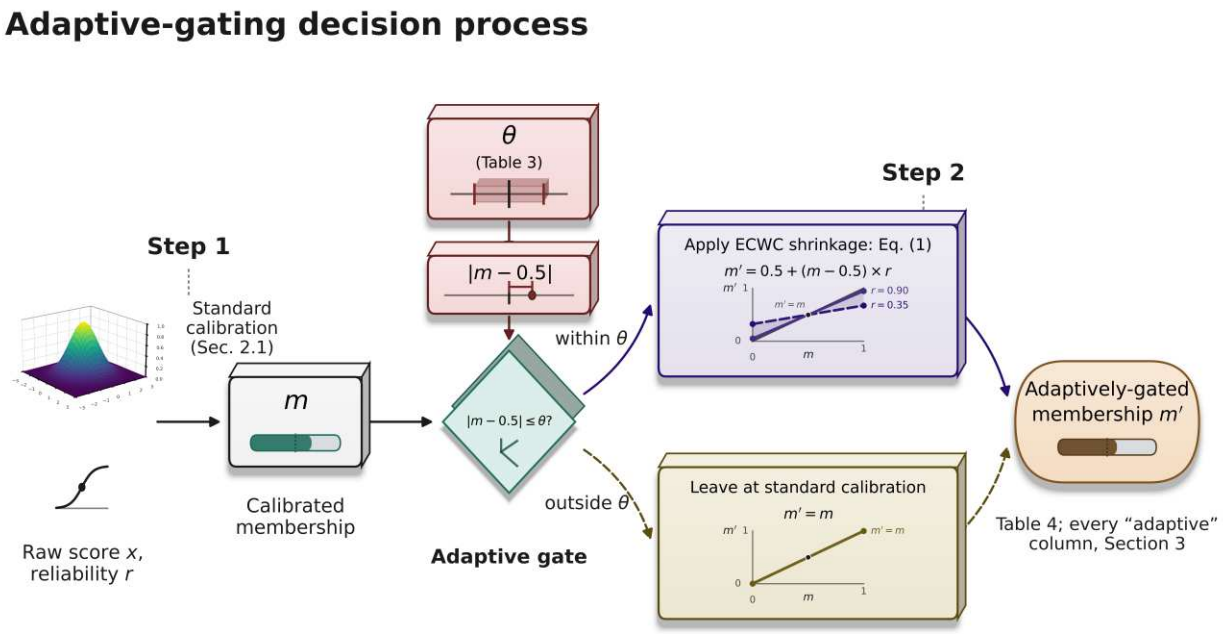


Fig. 2 The adaptive-gating decision process (Section 3.1, **Table 4**): a raw score is always standard-calibrated first (Section 2.1); only whether **Equation (1)**'s reliability-weighted shrinkage is then applied depends on whether the observed raw score falls within θ, that reliability level's own breakeven distance (**Table 3**), of the crossover

Figure 2 diagrams this two-step procedure exactly as Section 2.3 and **Table 4** apply it: the gate acts on the observed, not the true, raw score, using each reliability level's own published Gaussian breakeven distance as θ (**Table 3**). This single rule is what **Table 4**'s adaptive column, and every real-data validation's own adaptive-gating check in Sections 3.2–3.5, actually compute. The two branches are not evenly split even at the crossover (Section 3.1), which is why adaptive gating recovers most, but not all, of uniform ECWC's average-error cost (**Table 4**).

2.4. Real-Data Validation: Smart-Clothing Purchase Intention

Chen and Ye (2023) provide a hybrid PLS-SEM/fsQCA study of 225 Chinese consumers' intentions to purchase smart clothing: six conditions (functionality, expressiveness, aesthetics, perceived usefulness, perceived ease of use, attitudes) against purchase intention, each on a 5-point Likert scale with five items per construct, with Composite Reliability (CR) and Cronbach's alpha published in their own **Table 4**. I obtained the raw, de-identified item-level responses from the paper's public Supporting Information file and calibrated each construct's raw score (the mean of its five items) using the identical direct method as Section 2.1, with Chen and Ye's own anchors (full membership 5.0, crossover 3.5, full non-membership 1.0). Checking the raw item data against their published figures confirms four of six constructs and exposes a column-header transposition in the remaining two, an inconsistency this check exposes rather than resolves (full verification detail in Online Resource 1, Supplementary Note S15).

I use the necessity-consistent assignment throughout, since only that mapping bears on **Table 5**'s result, and each construct's published composite reliability as its ECWC weight, with Cronbach's alpha as a second check. I computed necessity consistency under both reliability sources and sufficiency consistency under composite reliability only, standard vs. ECWC-adjusted, plus a 10,000-replicate respondent-resampling bootstrap testing whether each shift direction reverses, and a jitter check perturbing each construct's reliability by independent uniform noise (±5%, ±10%, ±20%), 10,000 replicates per level, plus an adaptively gated version (Section 2.3's boundary) restricted to cases whose calibrated membership falls within a reliability-dependent band around the crossover.

2.5. Real-Data Validation: Fitness-App Continuance Intention

Zheng et al. (2026) provide an independent second technology-adoption test in a different domain: a hybrid PLS-SEM/Necessary Condition Analysis (NCA)/fsQCA study of 704 Chinese users' continuance intention to use mobile fitness apps, with five conditions (social attributes, entertainment attributes, service attributes, switching costs, satisfaction) against continuance intention, and composite reliability and Cronbach's alpha published in their own **Table 2**. I obtained the raw, de-identified item-level responses from the paper's public Mendeley Data deposit (CC BY 4.0) and fingerprint-matched them against the paper's own published per-construct statistics (full matching procedure and figures in Online Resource 1, Supplementary Note S15).

This study's calibration convention differs from every other real-data application here: full membership is set at each construct's own sample mean plus one standard deviation, crossover at the mean, non-membership at the mean minus one standard deviation, per the source paper's own methodology; ECWC's mechanism does not depend on which convention produced a calibrated membership. As in Section 2.4, I computed necessity consistency, standard vs. ECWC-adjusted, under both reliability sources, a matching bootstrap check, a jitter check (±5%, ±10%, ±20%, 10,000 replicates per level), and an adaptively gated version using the same boundary criterion as Section 2.3.

2.6. Real-Data Validation: Ragin's Breakdown-of-Democracy Benchmark

Sections 2.4 and 2.5 test ECWC within one domain, technology adoption; this section checks whether the boundary condition generalizes beyond it, applying ECWC to Ragin's Breakdown-of-Democracy dataset (Ragin 2008, Ch. 5, Table 5.2): calibrated membership scores for 18 interwar European countries on five conditions (Developed, Urban, Literate, Industrial, Stable) and one outcome (Breakdown of Democracy), using the values reproduced in Mendel and Korjani's technical report (2015, Table 6), the same source that report uses to validate its own competing calibration method.

Ragin published three calibration breakpoints for each of the six variables in this dataset (Ragin 2008). Mendel and Korjani (2015, Table 3) subsequently asked Ragin for his minimum- and maximum-uncertainty estimates around each breakpoint and published a left- and right-endpoint interval bounding its plausible range; I use these directly, combining each breakpoint's two published intervals into one span (left interval's lower bound to right interval's upper bound), giving three spans per variable, and deriving a reliability weight for each of the six variables as one minus the mean relative width of those spans (span width divided by the variable's total range), so narrower, more precise breakpoints receive a higher weight. This reuses uncertainty judgments Ragin himself provided and Mendel and Korjani (2015) already published, introducing no new assumption about reliability.

I then computed two standard, literature-consistent configurational tests: the necessity of instability (not-Stable) for democratic breakdown, and the sufficiency of instability combined with underdevelopment (not-Stable AND not-Developed) for breakdown, using the standard consistency formulas (Ragin 2008): first on the published calibrated memberships, then after ECWC adjustment using the weights derived above, then after an adaptively gated version using the same boundary criterion as Section 2.3. I also tested both shifts' sensitivity to each weight's estimation precision by perturbing the breakpoint-interval widths with independent elicitation noise (±5%, ±10%, ±20%, 10,000 replicates per level).

2.7. Real-Data Validation: Building Thermal-Comfort Data

Sections 2.4 and 2.5's weights are externally published; Section 2.6's is elicited directly from the original author. To illustrate ECWC where no such estimate exists, I apply it to the ASHRAE

Global Thermal Comfort Database II (Földváry Ličina et al. 2018): 109,033 field measurements from 809 buildings worldwide, each pairing an objective instrumented reading with a same-moment self-reported occupant vote. I aggregate to 130 buildings with at least 15 respondent records each for PMV (Predicted Mean Vote) and reported acceptability, calibrating PMV-predicted comfort against a crossover at ASHRAE Standard 55's ±0.5 PMV acceptability boundary (ANSI/ASHRAE 2023), applied to the negated |PMV| (PMV = 0 → full membership; |PMV| = 0.5 → crossover; |PMV| ≥ 1.5 → full non-membership); and reported acceptability against crossover and full-membership anchors at the standard's own 80% and stricter optional 90% thresholds (65% full-non-membership anchor, my own convention), using the same direct calibration method as Section 2.1.

Unlike Sections 2.4–2.6, no author-elicited or -published reliability estimate exists here, so I construct each building's weight from the data's own sampling precision: a DerSimonian–Laird (1986) method-of-moments estimate separates the between-building variance in mean PMV from each building's within-building sampling error, giving $r = \tau^2/(\tau^2+SE^2)$, the standard "reliability of a group mean" formula and a direct multi-group generalization of Kelley's (1927). Because this weight is self-constructed rather than elicited or published, I report it with correspondingly more qualified confidence. I tested necessity and sufficiency of PMV-predicted comfort for reported acceptability under standard, uniform, and adaptively gated ECWC (Section 2.3's boundary, applied per building), and repeated the gating check across five inclusion thresholds (10–30 minimum records per building).

2.8. Reproducibility: Code Availability

All calibration, simulation, and validation code was written in Python 3.11.15 (NumPy 2.4.4, SciPy 1.17.1, Matplotlib 3.10.9 for figures, pandas 3.0.2 for one table-construction script, and SymPy 1.14.0 for the Boolean-minimization check in Section 3.6), with seed = 42 for every reported point estimate (dedicated reseeding checks, Online Resource 1, vary it on purpose to test stability) and is available on request; the script package will be deposited on GitHub, archived to Zenodo, prior to publication (see Data availability). During preparation of this manuscript, the author used Claude Sonnet 5 (Anthropic) to help synthesize prior literature, implement the simulation, real-data validation, and benchmark-comparison code underlying the analyses reported here, and to draft portions of the manuscript text. The author independently verified every computational result reported in this manuscript and takes full responsibility for the content and accuracy of the published article.

3. Results

Section 3.1's boundary condition, not the ECWC correction it governs, is this paper's central finding, and Sections 3.2 through 3.5 test it against four real datasets.

3.1. When Reliability Weighting Helps, and When It Hurts

Table 2 shows, for a representative low-reliability case (r = 0.65), how far standard calibration and ECWC each land from the true, noise-free score: ECWC reduces this error only within about 10–15 points of the crossover (midpoint 50); beyond that range it pulls an already-accurate estimate toward artificial uncertainty.

Table 2. Mean absolute calibration error, standard vs. ECWC, by true raw score (r = 0.65, 20,000 replications per row). Benefit is computed from unrounded values, not from the three-decimal error columns shown.

True raw score (T)	|T − 50|	Standard error	ECWC error	Benefit (standard − ECWC)
10	40	0.020	0.177	−0.157 (hurts)
25	25	0.068	0.170	−0.101 (hurts)
35	15	0.126	0.139	−0.013 (hurts)
40	10	0.154	0.122	+0.032 (helps)
50	0	0.182	0.118	+0.064 (helps)
60	10	0.152	0.120	+0.032 (helps)
65	15	0.124	0.138	−0.014 (hurts)
75	25	0.068	0.169	−0.101 (hurts)
90	40	0.019	0.176	−0.157 (hurts)

Table 3 extends this to five reliability levels (r = 0.35–0.90) under Gaussian noise, and four further noise patterns at r = 0.65 (uniform; skew-normal, shape parameter = ±4; Student's t, df = 3). Across all nine conditions the breakeven distance stays within a 10.5–14.6-point band: the two skew-normal conditions match Gaussian almost exactly, while the heavy-tailed condition contracts to 11.3 points (95% CI [11.3, 11.4], no overlap with Gaussian's 13.5). The crossover benefit itself shrinks steadily as reliability rises, from +0.179 at r = 0.35 to +0.006 at r = 0.90. Online Resource 1 tests every reliability/noise combination jointly rather than one at a time.

Table 3. Breakeven distance from the crossover at which ECWC's benefit turns to a cost, varying reliability and noise model one at a time (marginally, not jointly): five reliability levels under Gaussian noise, then four further noise models at a single fixed reliability (r = 0.65), nine conditions total, not the full 5 × 5 grid (see Online Resource 1; 20,000 replications per true-raw-score value; bracketed values are 95% confidence intervals, Monte Carlo-propagated from each point's sampling error).

Reliability (r)	Noise model	Breakeven distance [95% CI]	Benefit @ crossover	Benefit @ |T−50|=40
0.35	Gaussian	13.5 [13.5, 13.6]	+0.179	−0.273
0.50	Gaussian	13.6 [13.5, 13.6]	+0.117	−0.219
0.65	Gaussian	13.5 [13.4, 13.5]	+0.064	−0.157
0.80	Gaussian	12.5 [12.4, 12.6]	+0.023	−0.090
0.90	Gaussian	10.5 [10.4, 10.7]	+0.006	−0.045
0.65	Uniform	14.6 [14.5, 14.6]	+0.071	−0.156
0.65	Skew-normal (α = +4)	13.5 [13.4, 13.5]	+0.064	−0.157
0.65	Skew-normal (α = −4)	13.5 [13.4, 13.5]	+0.063	−0.157
0.65	Student's t (df = 3)	11.3 [11.3, 11.4]	+0.049	−0.159

Every breakeven-distance CI is no wider than about ±0.2 points, far tighter than the roughly 3-point spread across reliability levels, confirming the trend is genuine rather than Monte Carlo noise. Section 2.3's analytical (MSE-based) breakeven distance tracks the exact numerically-integrated answer closely, though not exactly, across calibration functions and error criteria (Online Resource 1). The correlated multi-condition check (Section 2.3) validates cleanly, reproducing **Table 4**'s single-condition numbers almost exactly under fully shared noise, with uniform ECWC's cost relative to standard calibration holding regardless of correlation. Reliability weighting is thus a targeted correction: its benefit concentrates in genuinely ambiguous, low-reliability cases, and its cost falls on confidently-extreme ones.

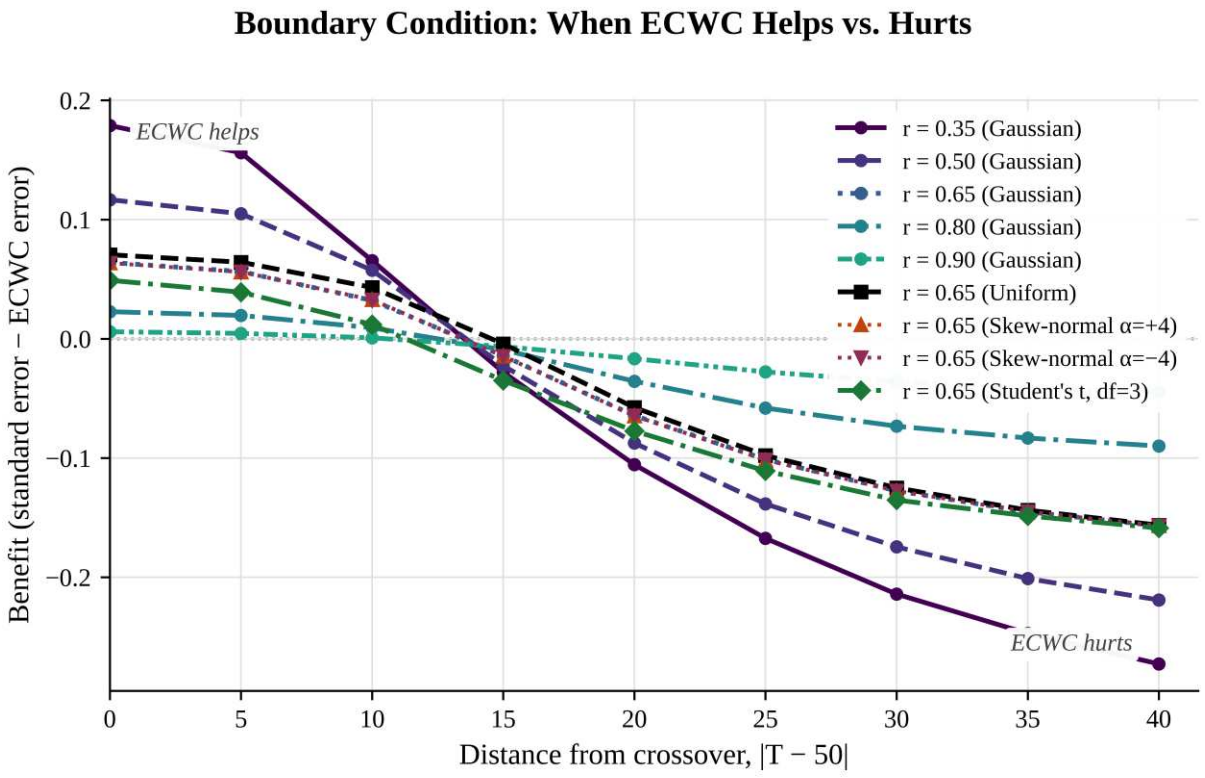


Fig. 3 Benefit (standard error − ECWC error) as a function of distance from the crossover, |T − 50|, across five reliability levels under Gaussian noise and, at r = 0.65, under four further variance-matched noise models (uniform; skew-normal with shape parameter = +4; skew-normal with shape parameter = −4; Student's t with df = 3). Positive values indicate ECWC reduces calibration error; negative values indicate it increases it. Underlying data: 20,000 replications per point (**Table 3**)

Figure 3 plots the full benefit curve behind **Table 3**: all nine conditions cross zero within the same 10.5–14.6-point band, with the skew-normal curves tracking Gaussian and uniform closely while the heavy-tailed curve crosses zero earlier. Benefit near the crossover grows steeply as reliability falls while cost at the sweep's far end grows more gradually, and varying noise model at fixed reliability narrows both effects considerably. Online Resource 1 recasts this Gaussian-noise data as a decision map: its zero-benefit boundary reproduces **Table 3**'s breakeven distances, nearly flat through r = 0.65 before narrowing at higher reliability (12.5 at r = 0.80, 10.5 at r = 0.90) and further still under heavy-tailed noise, so the map applies to the Gaussian case only.

Table 3 and **Figure 3** vary reliability and noise pattern one at a time; Online Resource 1 tests the full 5 × 5 grid jointly (all 25 combinations), confirming high reliability and heavy-tailed noise compound rather than offset (narrowest cell 7.8 points, vs. 10.5–11.3 for either factor alone) and validating closely against **Table 3**'s marginal values, within 0.1 point at every reliability level tested. The boundary is therefore better stated as roughly 8–15 points, narrowing further to about 8 when high reliability and heavy-tailed noise combine. This assumes the true score is known, but only the noisy observed score is available in practice; Section 4.1's fix, gating shrinkage on the

observed score's own distance from the crossover rather than applying it uniformly, is tested directly here using each reliability level's Gaussian breakeven distance (**Table 3**) as the gating threshold.

Table 4 reports mean absolute calibration error across the full T = 10–90 sweep, for standard calibration, uniform ECWC, and this adaptive gated version. Averaged across the full range, uniform ECWC's error exceeds standard calibration's at every reliability level, since the sweep is dominated by cases outside the narrow helps-region; adaptive gating recovers 95.9–97.2% of that cost at every level tested. At r = 0.65, mean absolute error falls from 0.152 (uniform ECWC) to 0.097 (adaptive) against a 0.095 standard baseline, capturing most but not all of the crossover benefit (+0.039 vs. +0.064 at T = 50), since the gate acts on the noisy observed, not true, score. This recovery is not uniform: across most of the helps-region itself, adaptive gating's own raw error exceeds uniform ECWC's, though both still beat standard calibration, because only about 80% of observed draws at the crossover fall inside the gate, a gap that narrows as reliability rises (99.9% gated-in at r = 0.90). Online Resource 1 tests the gate's design under wider and softer thresholds.

Table 4. Mean absolute calibration error under standard calibration, uniform ECWC, and adaptive (observed-score-gated) ECWC, by reliability level, averaged across the full T = 10–90 sweep (20,000 replications per point, Gaussian noise). θ is the gating threshold: each reliability level's own already-published Gaussian breakeven distance (**Table 3**). "Harm recovered" is the share of uniform ECWC's average-error excess over standard calibration that adaptive gating eliminates, computed from unrounded values, not from the three-decimal error columns shown.

Reliability (r)	θ (points)	Standard error	Uniform ECWC error	Adaptive ECWC error	Harm recovered
0.35	13.5	0.168	0.253	0.171	96.7%
0.50	13.6	0.133	0.206	0.136	96.3%
0.65	13.5	0.095	0.152	0.097	95.9%
0.80	12.5	0.054	0.091	0.056	97.1%
0.90	10.5	0.027	0.047	0.028	97.2%

Table 4 settles the practical question: uniform ECWC is, on average, worse than standard calibration at every reliability level tested, while adaptive gating recovers the great majority of that cost, so I recommend the gated version for practical use. Uniform ECWC is still retained in Sections 3.2–3.5 for continuity, since gating requires first deriving a dataset- and reliability-specific boundary (Section 2.3); each real-data section reports both and identifies which it treats as primary.

3.2. A First Real-Data Application: Technology-Adoption Purchase Intention

ECWC's motivating claim is that it applies wherever fsQCA draws on heterogeneous, unequally reliable sources, not only the political-science benchmark validated below (Section 3.4). Chen and Ye's (2023) smart-clothing purchase-intention study (dataset, calibration, and data-verification

procedure detailed in Section 2.4) provides a direct test, calibrating six conditions against purchase intention (named in **Table 5**) using each construct's published Composite Reliability and Cronbach's alpha as the ECWC weight.

Table 5 applies ECWC using composite reliability as the weight for each condition. All six weights fall in a genuinely mid-range band (0.872–0.931), comparable to Ragin's own weights (Section 3.4: 0.712–0.883), so the correction moves every condition here, not just a few outliers. Necessity falls modestly for all six conditions (−0.003 to −0.011) while sufficiency rises modestly (+0.006 to +0.017), opposite the Ragin and ASHRAE validations below, expected since ECWC's shift direction follows each dataset's own raw-score distribution relative to the crossover.

Table 5. Necessity and sufficiency consistency, standard vs. ECWC-adjusted, on Chen and Ye's smart-clothing purchase-intention sample (N = 225), using Composite Reliability (CR) and Cronbach's alpha (α) as the weight for each of six conditions (Chen and Ye 2023, Table 4). Necessity is reported under both reliability sources; sufficiency, in this validation, under composite reliability only. Both Δ columns are CR-based: the necessity Δ is CR-based ECWC minus Std., skipping the intervening α-based column shown for comparison; the sufficiency Δ is CR-based ECWC minus Std., with no intervening column, immediately to its left. Δ values are computed from unrounded values, not from the three-decimal columns shown.

Condition	Necessity				Sufficiency		
	Std.	ECWC (CR)	ECWC (Cronbach α)	Δ (CR)	Std.	ECWC	Δ
Functionality	0.910	0.900	0.895	−0.010	0.778	0.795	+0.017
Expressiveness	0.904	0.896	0.893	−0.008	0.793	0.805	+0.012
Aesthetics	0.825	0.822	0.821	−0.003	0.885	0.892	+0.007
Perceived usefulness	0.901	0.894	0.892	−0.007	0.868	0.874	+0.006
Perceived ease of use	0.878	0.874	0.872	−0.005	0.855	0.861	+0.006
Attitudes	0.952	0.941	0.937	−0.011	0.861	0.867	+0.007

The more consequential result is a potential classification change. Chen and Ye's own criterion for a necessary condition is necessity consistency above 0.90 (Chen and Ye 2023); four conditions clear that bar under standard calibration, but only attitudes remains comfortably above it under ECWC (**Table 5**), at face value a four-to-one reclassification, though functionality's ECWC necessity (precisely 0.8995) sits only half a rounding unit below 0.90. Bootstrap resampling shows this is a genuinely mixed picture, not a clean reclassification, though shift direction itself is confirmed with near-certainty for five of six conditions and strongly for the sixth (Online Resource 1).

A jitter check mirroring Section 3.4's (±20%, 10,000 replicates; Online Resource 1) leaves every condition's shift sign unchanged at every level (P(sign flips) < 0.0001): unlike Ragin's benchmark, these shifts are not fragile to reliability-estimation imprecision. Adaptive gating shrinks every condition's already-small shift here (0.003–0.017 in **Table 5**'s CR-based columns) to within ±0.007 of zero, with 44–76% of respondents inside the ambiguity band, and unlike Ragin's benchmark

does not cleanly reverse sign here, only magnitude: all three necessity reclassifications return above 0.90 under composite reliability, but only one under Cronbach's alpha. Gated shifts are treated as primary.

This validation is bounded to one product category (smart clothing) and one national context (China, N = 225); it is the manuscript's most directly comparable application to Ragin's (Section 3.4), both using externally published, per-variable reliability statistics and both showing a differentiated effect at the classification level alongside a robust one at the directional level.

3.3. A Second Real-Data Application: Technology-Adoption Continuance Intention

A single technology-adoption validation cannot show whether Section 3.2's pattern generalizes. Zheng et al. (2026) provide an independent second test: a mobile fitness-app continuance-intention study (dataset, construct-mapping, and sample-relative calibration convention detailed in Section 2.5), with five conditions against continuance intention as the outcome (named in **Table 6**), using each construct's published composite reliability and Cronbach's alpha as the ECWC weight.

Table 6 reports necessity consistency for each condition, standard vs. ECWC-adjusted. Under standard calibration, no condition clears a conventional 0.90 threshold; satisfaction comes closest (0.899). Under ECWC it crosses (0.904), the opposite direction from Section 3.2's finding. The other four conditions move by similarly small amounts (−0.009 to +0.008) and none approaches the threshold: a smaller effect than the Ragin or smart-clothing validations, and a second, independent confirmation that ECWC behaves sensibly on real technology-adoption data rather than a second threshold-crossing story.

Table 6. Necessity consistency, standard vs. ECWC-adjusted, on Zheng et al.'s mobile fitness-app continuance-intention sample (N = 704), using Composite Reliability (CR) and Cronbach's alpha (α) as the weight for each of five conditions (Zheng et al. 2026, Table 2).

Condition	Standard	ECWC (CR)	ECWC (Cronbach α)
Social attributes	0.821	0.829	0.829
Entertainment attributes	0.873	0.877	0.877
Service attributes	0.856	0.854	0.853
Satisfaction	0.899	0.904	0.904
Switching costs	0.850	0.844	0.841

Bootstrap resampling confirms the shift direction for four of five conditions, weaker for the fifth; a complementary jitter check shows these already-small shifts are the most reliability-sensitive of the manuscript's two technology-adoption applications (Online Resource 1).

Adaptive gating shrinks four of the five conditions' already-small uniform shifts further toward zero, while the fifth (switching costs) flips sign, the condition with by far the fewest respondents gated in (Online Resource 1). This validation is bounded to a single national context and product

domain; its value lies in combination with Section 3.2's, two independent datasets both showing small, directionally robust, bounded shifts.

3.4. Generalizability Check: Ragin's Breakdown-of-Democracy Benchmark

I apply ECWC here as a generalizability check: this benchmark spans a different substantive domain (political science, not technology adoption) and a different, author-published calibration-uncertainty convention from Sections 3.2 and 3.3, and it is the same real-data validation Mendel and Korjani (2015, 2018) use to benchmark the closest prior competing calibration method. **Table 7** reports the resulting reliability weights for each of the six Breakdown-of-Democracy variables.

Table 7. Reliability weights derived from Ragin's own breakpoint-uncertainty estimates, elicited from him and published in Mendel and Korjani (2015, Table 3).

Variable	Reliability weight (r)
A: Developed	0.746
B: Urban	0.874
C: Literate	0.712
D: Industrial	0.808
E: Stable	0.883
O: Breakdown of Democracy	0.843

Necessity and sufficiency, standard vs. ECWC-adjusted, shift modestly in the same, historically sensible direction but differ in strength (**Table 8**): sufficiency of (not-Stable AND not-Developed) stays comfortably above the conventional 0.8 threshold both before and after, while necessity of not-Stable remains well below it even after the shift, consistent with instability being one contributing condition among several rather than, alone, necessary for democratic breakdown. Online Resource 1 breaks down each country's contribution: ECWC shrinks Portugal, Poland, and Spain (the highest standard-calibration contributors) the most, while the more historically contested trajectories (Germany, Austria, France) gain influence, a redistribution Section 3.1's boundary condition addresses below.

Online Resource 1 visualizes this shift by country (Efron and Hastie's 2016 paired before/after format): shift direction is fixed entirely by which side of the crossover a country's standard contribution falls on, with the five above 0.5 shrinking down and the thirteen below shrinking up, though magnitude does not simply track distance from the crossover. A robustness check recomputing the six weights with a quadratic rather than linear penalty on Ragin's published interval widths leaves direction and per-country pattern unchanged, though magnitude shrinks substantially under the steeper penalty: the redistribution is not an artifact of the weight formula, only its exact size (Online Resource 1).

Every shift reported so far in this section uses uniform ECWC. Section 3.1's boundary condition, however, says shrinkage only reduces error within roughly 8–15 points of the crossover; Section

4.1 names the resulting adaptive, gated version, which **Table 4** shows recovers 95.9–97.2% of uniform ECWC's average-error cost in simulation. Sections 3.2 and 3.3 already apply it; this section applies it for the first time to Ragin's own expert-calibrated benchmark, the validation closest to the boundary condition's assumptions. Because every calibration function here maps the full-membership anchor to 0.953 and the crossover to 0.5, **Table 3**'s raw-score breakeven distances convert exactly into a calibrated-membership gate: a case is shrunk if its membership falls within a reliability-dependent band around 0.5 (each variable's own weight, **Table 7**), left at standard calibration otherwise; three alternative boundary derivations agree to the last reported digit.

The result reverses the uniform-ECWC finding. Between 5.6% and 27.8% of the 18 countries are gated in per variable: most of Ragin's own calibrated memberships already sit confidently away from the crossover, exactly where Section 3.1 says shrinkage should not apply. Portugal, Poland, and Spain, the three countries uniform ECWC shrinks most, are gated out on every variable relevant to the sufficiency test, so their contribution stays at standard calibration (0.950, 0.880, 0.800). **Table 8** reports the resulting shifts: necessity moves from +0.039 under uniform ECWC to −0.004 under adaptive gating; sufficiency moves from +0.035 to −0.003, both the reverse sign of what uniform ECWC reported, and both close to zero.

Table 8. Necessity and sufficiency consistency on the Breakdown-of-Democracy data under standard calibration, uniform ECWC, and adaptively-gated ECWC (Section 3.4 methodology above). Shift columns are computed from unrounded values, not from the three-decimal columns shown.

Test	Standard	Uniform	Adaptive	Shift (unif.)	Shift (adapt.)
Necessity of not-Stable for Breakdown	0.656	0.696	0.653	+0.039	−0.004
Sufficiency of (not-Stable AND not-Developed) for Breakdown	0.893	0.928	0.889	+0.035	−0.003

This does not show ECWC is wrong: real cases have no known ground truth, so it verifies only that gating changes which countries get shrunk, not that it reduces error the way **Table 4** does. It does establish that the uniform-ECWC redistribution reported above (away from Portugal, Poland, and Spain) is not what a principled correction should do, since those three are exactly the confidently-extreme cases the boundary condition says shrinkage should not touch: ECWC's role is to leave most of Ragin's judgments untouched, acting only on the genuinely marginal minority. The same gated check on Chen and Ye's smart-clothing data (Section 3.2) does not reverse as cleanly, only the shrinkage-toward-zero pattern holding stable. Per Section 3.1's recommendation, **Table 8**'s gated shifts, not its uniform ones, are this section's primary finding.

Each variable's weight is an estimate from only three published breakpoint spans (Section 2.6); perturbing each span's width by independent uniform elicitation noise (±5%, ±10%, ±20%; 10,000 replicates; Online Resource 1) leaves both shifts strictly positive at every level, and neither point estimate ever crosses the conventional 0.80 threshold under the same jitter (Online Resource 1). This check was run on the uniform-ECWC point estimates, not the gated ones identified above as primary.

3.5. A Third, Illustrative Application: Building Thermal-Comfort Data

To illustrate ECWC where no such estimate exists, I apply it to the ASHRAE Global Thermal Comfort Database II (Földváry Ličina et al. 2018): 109,033 field measurements from 809 buildings worldwide (Section 2.7). PMV, the field's standard calculated comfort index, predicts actual reported thermal sensation only imperfectly across building types (Cheung et al. 2019), the mixed-reliability situation this paper's motivating examples describe (Section 1). I aggregate to 130 buildings with at least 15 respondent records each and test necessity and sufficiency of PMV-predicted comfort for reported acceptability, calibrated as in Section 2.7: negated PMV, with anchors reordered accordingly ($0 \rightarrow 0.953$; $0.5 \rightarrow 0.5$; $1.5 \rightarrow 0.047$). The reliability weight is self-constructed from the data's own sampling precision ($\tau^2 = 0.137$; Section 2.7), reported with correspondingly more qualified confidence.

Table 9 reports necessity and sufficiency consistency under standard calibration, uniform ECWC, and adaptively-gated ECWC. Under standard calibration, PMV-predicted comfort is moderately necessary and sufficient for reported acceptability (0.668 and 0.699), neither clearing the conventional 0.8 threshold, consistent with PMV predicting reported comfort only imperfectly (Cheung et al. 2019). Reliability weights are high throughout: only 5 of 130 buildings fall below 0.85, so uniform ECWC leaves most buildings nearly unchanged and its shift is small (roughly 0.001–0.002), an order of magnitude smaller than Sections 3.2–3.4's shifts.

Table 9. Necessity and sufficiency consistency on the ASHRAE thermal-comfort building sample (N = 130, $\tau^2 = 0.137$, reliability weights ranging 0.690–0.999, mean 0.956), under standard calibration, uniform ECWC, and adaptively-gated ECWC (Section 3.5 methodology above).

Test	Standard	Uniform	Adaptive	Shift (unif.)	Shift (adapt.)
Necessity of PMV-predicted comfort for reported acceptability	0.6675	0.6680	0.6696	+0.0005	+0.0021
Sufficiency of PMV-predicted comfort for reported acceptability	0.6986	0.7002	0.6992	+0.0016	+0.0006

The same adaptively-gated check from Sections 3.2–3.4 applies here per building rather than per variable. At the reported 15-record threshold, 51.5% of buildings are gated in, and the gated shift keeps the same sign as the uniform shift, unlike Section 3.4's benchmark, where the two disagree; Online Resource 1 confirms the gated shift's sign holds positive across all five inclusion thresholds tested, though the uniform shift itself flips sign at the loosest one (10 records). Across all four validations, gating shrinks most shifts toward zero in Sections 3.2 and 3.3 without a clean overall reversal, reverses the shift cleanly in Section 3.4's benchmark, and, uniquely, holds its sign stably here; Section 4.1 draws out what that pattern means for when the correction should be trusted.

3.6. Boolean-Minimization Sensitivity on an Independent Benchmark

Sections 3.2–3.5 test whether ECWC's correction shifts necessity and sufficiency consistency; none tests the further, configurational question of whether a shift this size is large enough to change

which combinations of conditions a full Boolean minimization reports as sufficient (Section 4.5). This section reports a direct test of that question on an external, independently published benchmark, not a fifth validation of ECWC itself, since it perturbs raw calibration disagreement rather than ECWC-adjusted memberships. Using Ragin's own original calibration for the Breakdown-of-Democracy data (2008, Ch. 5, Table 5.2), not the Mendel-and-Korjani (2015) reproduction behind Section 3.4's "Stable" condition, I construct a truth table and minimize it (Quine–McCluskey) exactly as Ragin describes. The point estimate exactly reproduces his published solution, consistency (0.871), and coverage (0.790), confirmed by an automated equality check against his reported values (full minimization parameters and solution formula in Online Resource 1, Supplementary Note S13): an external validation of the minimization procedure itself, independent of anything else this paper claims.

Two schemes then perturbed this validated point estimate and re-minimized under the identical convention (construction detail and full match-rate figures in Online Resource 1, Supplementary Note S13): independent noise on each case's membership score, and noise applied in log-odds space, which compresses perturbation near the anchors and expands it near the crossover, the more defensible model of real analyst disagreement (**Figure 4**). Both agree qualitatively: several of Ragin's own truth-table rows sit within 0.83–0.91 consistency on only one or two cases each, so ordinary calibration disagreement is frequently enough to move them across the 0.80 cutoff and change the minimized solution, establishing that the configurational vulnerability Section 4.5 raises is real, not hypothetical, though not whether ECWC's own correction would stabilize it (Section 4.5).

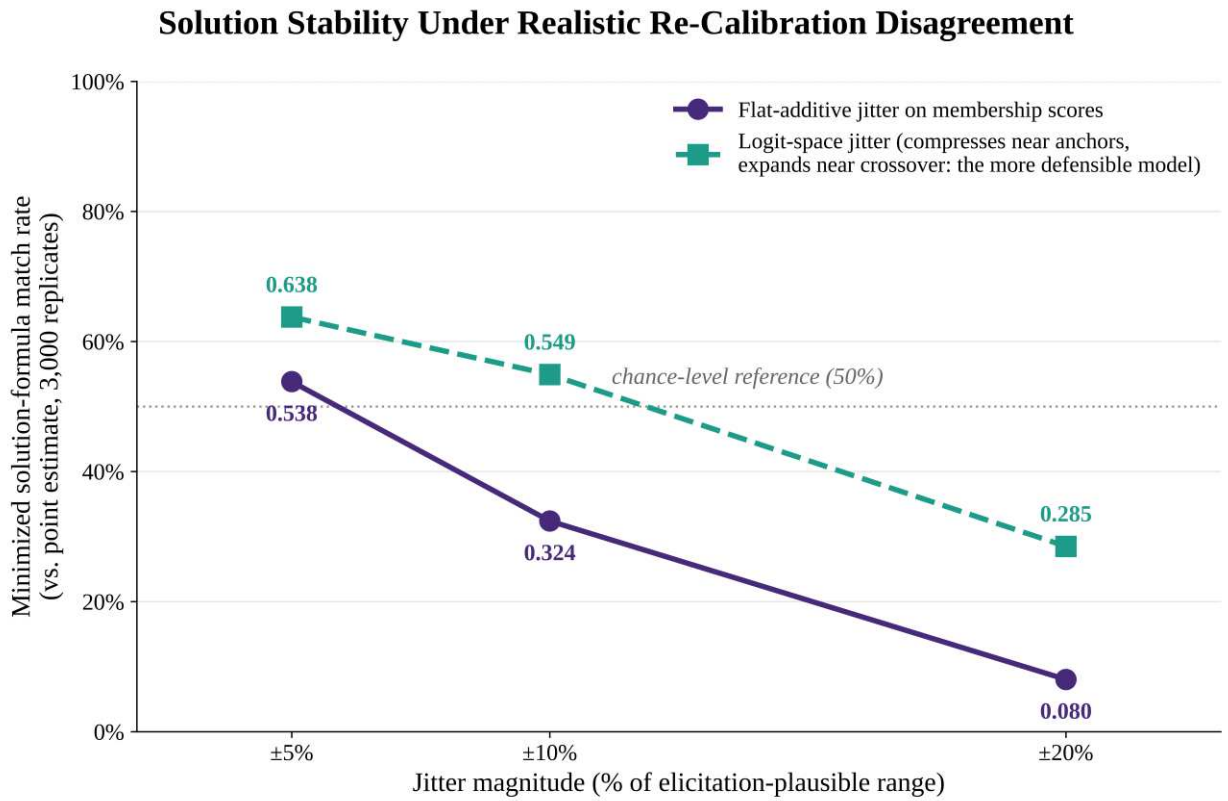


Fig. 4 Minimized solution-formula match rate against the point estimate (3,000 replicates per level) as re-calibration disagreement grows, for both perturbation schemes described above, on Ragin's own Breakdown-of-Democracy benchmark. The logit-space scheme (compressing perturbation near the calibration anchors, expanding it near the crossover) is the more defensible model of real disagreement between analysts, and is still less than a coin flip at ±20% jitter

Figure 4, read against the chance-level reference line, shows the two schemes diverging in what they imply is practically safe. At ±10% jitter, the logit-space scheme's minimized solution is still more often right than wrong (0.549), while the flat-additive scheme has already fallen below a coin flip (0.324); the logit-space scheme is not overtaken by chance until somewhere between ±10% and ±20%, and even at ±20% remains 3.5 times as likely to match the true solution as the flat-additive scheme (0.285 vs. 0.080). Since the logit-space scheme is also the more defensible model of real analyst disagreement (above), this comparison if anything understates the minimized-solution stability an analyst should expect to lose under ordinary calibration disagreement, reinforcing the configurational vulnerability Section 4.5 discusses.

4. Discussion

The results above test the boundary condition through simulation (Section 3.1), four real-data validations (Sections 3.2–3.5), and a Boolean-minimization robustness check on an independent benchmark (Section 3.6); this section interprets what that record means for practice (4.1), situates the real-data shifts against existing calibration-robustness benchmarks (4.2), positions the

correction relative to fsQCA's own calibration epistemology (4.3), extends the argument to the broader research this paper's motivation spans (4.4), and states this work's limitations directly (4.5).

4.1. A Reliability-Validity Tradeoff, Not Merely a Targeted Correction

The boundary condition in Section 3.1 is the paper's most consequential finding: shrinking toward maximal ambiguity helps only when a low-confidence estimate's true value is ambiguous, and hurts when that value is confidently extreme, a distinction no uniform, source-level weight can make in advance. This is the same principle behind James–Stein shrinkage (Efron and Hastie 2016), Kelley's formula (Section 2.2), and an analog Willink (2026) derived for metrology. Section 2.3's closed-form adaptation stays within 3.0–4.0% of the exact boundary for the logistic and probit cases, though it is looser for complementary log-log under the MSE criterion (within 16.7%), a gap the skewness-corrected MAE criterion narrows unevenly across reliability levels (Online Resource 1). This is not only a simulation artifact: gating Ragin's real, calibrated memberships (Section 3.4) reverses the sign of the uniform-ECWC shift otherwise reported for that dataset (**Table 8**), because the countries uniform ECWC shrinks most (Portugal, Poland, Spain) are exactly the confidently-extreme cases the boundary condition says a correction should leave alone.

This is a formal instance of an older, previously only qualitative warning in measurement theory: reliability and validity are distinct, and correcting for one does not guarantee, and can cost, the other (Cronbach and Meehl 1955); Section 3.1's boundary condition gives it an exact, closed-form location within a bounded, anchor-based calibration procedure for the first time. The same cost-region pattern recurs elsewhere: factor scores in structural equation models (Burghgraeve et al. 2021), regression-to-the-mean corrections in ecological data unreliable enough that not correcting is sometimes preferable (Fontanari and Santos 2026a, 2026b, the same conclusion **Table 4** reaches here), Empirical Bayes shrinkage in regression (Chen et al. 2025), and survey calibration, where marginal validity does not guarantee subgroup-level reliability (Rafe and Das 2026; Online Resource 1). None addresses fsQCA's own bounded, nonlinear calibration function directly, but their recurrence suggests this tradeoff is a general property of reliability-weighted correction, not one method's idiosyncrasy.

This is the paper's practical recommendation: apply the adaptively gated version of ECWC, not the uniform version, wherever the added derivation step (Section 2.3) is feasible. Gating does not fully close the gap, recovering most but not all of uniform ECWC's simulated cost and still underperforming it inside the helps-region itself (Section 4.5), but a case gated near the boundary is where the correction and underlying calibration are least certain, so its value lies as much in flagging that uncertainty as in the point estimate itself. ECWC also addresses a different vulnerability than threshold-sweep tools such as ThSQCA (Toyoda 2026) or the robustness-test protocol (Oana and Schneider 2024), which test sensitivity to threshold choice rather than weighting a fixed threshold's output by reliability: reimplementing ThSQCA's rule on the same

Ragin data moved necessity and sufficiency several times more than ECWC's own shifts (see Data availability).

4.2. Contextualizing the Real-Data Result

On the Breakdown-of-Democracy benchmark, uniform ECWC's modest shifts preserve both tests' conclusions while gating reverses their sign, since most of Ragin's countries sit confidently away from the crossover (Section 3.4, **Table 8**); neither overturns Ragin's classic analysis, though the reversal means uniform ECWC is not the more defensible number. Against two independent robustness benchmarks (Cooper and Glaesser 2016; Mendel and Korjani 2015, 2018), ECWC's shifts and jitter-tolerance sit within a range neither would consider remarkable, so neither result appears fragile to the estimation uncertainty a reviewer would raise (Online Resource 1). Sections 3.2, 3.3, and 3.5 extend this picture to three further domains, though none has a comparable external benchmark to situate against.

4.3. Calibration as Theoretical Judgment, Not Statistical Estimation

Positioning ECWC requires locating it between two literatures that, until now, have not met: Hug (2013) and Maggetti and Levi-Faur (2013) diagnosed QCA's vulnerability to measurement error without resolving it (Section 1), and the sensitivity-diagnostics tools that followed (Skaaning 2011; Krogslund et al. 2015; Oana and Schneider 2024; Toyoda 2026) test only how much a computed solution would change under hypothetical coding error, after calibration is already finished. None asks whether calibration itself, the step at which Ragin's (2008, Ch. 5) own theoretical account says membership should already reflect a considered, case-based judgment, should be adjusted for a source's known reliability before that judgment is put to use. ECWC is the first calibration procedure to do so.

Ragin (2008, Ch. 5) argues that calibration anchors and the calibration function should be set from substantive, case-based theoretical knowledge, not derived from a variable's own distribution; a case's set membership is a theoretical claim about what that case is, not a statistical estimate of an underlying continuous quantity. ECWC does not revisit any of this: the anchors and calibration function remain exactly as the researcher set them, applied only after producing a case's calibrated value. What ECWC treats as uncertain is that calibrated value itself, in the sense classical test theory already distinguishes an observed score from a true one (Section 2.2): a case's calibrated membership is at best an estimate of what the same function would assign to that case's true raw value, and a low-confidence source is one where the recorded and true raw values are more likely to diverge. **Equation (1)** shrinks that membership toward the crossover in proportion to source-level confidence, not because the calibration function's theoretical content is in doubt, but because a low-confidence case's output is a less trustworthy stand-in for that estimate.

Because this shrinkage acts on calibrated membership before truth-table construction, a case pulled toward the crossover under low reliability cannot change which side of the crossover it falls on (Section 4.5), so its coarse row assignment is unaffected; what it can still change is that row's fuzzy

consistency score, and therefore whether the row clears the threshold used to code its outcome, a downstream, configurational consequence this paper's checks do not test directly (Section 4.5) and a real remaining limitation. ECWC is best understood not as a departure from fsQCA's calibration tradition but as an extension of a step that tradition has always presupposed: a case's calibrated membership is only as trustworthy as the raw score it was computed from, which can be wrong in a way the theory-driven function cannot detect on its own.

4.4. Cross-Domain Relevance

The broader problem, unequally reliable data within a single fsQCA analysis, is not domain-specific: fsQCA's spread from comparative political science (Ragin 2008) into technology-adoption, business/management, and sustainability research (Section 1) already spans this range (Li et al. 2025b; Abbasi et al. 2022; Kumar et al. 2022). This paper's four validations are the fit argument: the same reliability problem recurs across fsQCA's uses, from technology adoption's mixed self-reported and system-logged data to nature exposure's wider measurement heterogeneity (Online Resource 1). Section 2.6's cross-national validation fits the same comparative structure as a 20-country calibration framework on health data (Batista Hernandez et al. 2026). **Equation (1)** is domain-agnostic, needing only a bounded reliability estimate (composite reliability, Cronbach's alpha, or a DerSimonian–Laird weight, Section 2.7); **Equation (2)**'s boundary condition, however, applies only to that kind of estimate, not a differently-sourced signal such as a model's own output confidence. ECWC offers a low-cost way to make this assumption explicit and testable across all of them (Oana and Schneider 2024).

The two technology-adoption applications (Sections 3.2, 3.3) illustrate this from an adoption-cost angle: composite reliability and Cronbach's alpha, the statistics ECWC needs, are already routine output of the hybrid PLS-SEM/fsQCA design both source studies use (Chen and Ye 2023; Zheng et al. 2026), so for a researcher already running that design, ECWC's weight comes essentially free. This does not generalize: the 34-study audit behind **Table 1** spans several fields but not technology adoption specifically, so how commonly technology-adoption fsQCA work pairs with a prior PLS-SEM stage remains open, and the same reporting gap applies where it does not.

4.5. Limitations

This section's limitations fall into two groups: how well-established the boundary condition and gated correction are, and how far the validations and literature audit generalize. The boundary narrows to roughly 8–15 points across the full joint grid of reliability levels and noise models, tightening further under heavier-tailed noise (Online Resource 1), and the analytical (MSE-based) boundary targets a different error criterion than the simulation's MAE, with each criterion's accuracy characterized separately across calibration functions (Online Resource 1). Adaptive gating was previously validated only in simulation; extending it to all four real-data validations shows only that gating changes which cases are shrunk, not that it reduces error against an unknown true score, and results are mixed across the four datasets (Sections 3.2–3.5), consistent with the boundary condition. Even where validated, gating is not uniformly better: in simulation it

underperforms uniform ECWC across most of the helps-region, since gating requires a dataset- and reliability-specific boundary the uniform version does not, a real qualification of this paper's no-new-data framing.

Three of the four real-data applications weight by variable or construct, not case (Section 3.4, **Table 7**; Sections 3.2–3.3, **Tables 5, 6**); the fourth, ASHRAE (Section 3.5), weights case by case (**Table 9**); a systematic comparison remains for future work. Every check above stops at necessity/sufficiency consistency scores; none propagates ECWC-adjusted memberships through full truth-table Boolean minimization, the further step Korjani and Mendel's interval type-2 fsQCA framework addresses differently (Korjani and Mendel 2014). One property of **Equation (1)** partially bounds this risk (Section 4.3): ECWC can never move a case's membership across the crossover, so truth-table row assignment is invariant to the correction, though each row's fuzzy consistency score can still shift enough to change the minimization threshold decision (tested in **Table 5**, Section 3.2, and Section 3.6). The three real-data reliability weights also depart from **Equation (2)**'s noise model differently, and none is tested against the boundary condition directly (full algebraic detail and noise-model comparison in Online Resource 1, Supplementary Note S18).

Each real-data validation is further bounded in scope (by dataset, elicitation source, or calibration convention), detailed in Online Resource 1, as are the Introduction's two audit rounds, convenience rather than pre-registered samples skewing China-context and open-access. The audit itself was single-coder, so no formal inter-rater reliability statistic can be reported for it either, the same omission this paper documents as a field-wide pattern; a subsample cross-check by a second coder is left for future work. Public health was among the audited domains (**Table 1**); nature exposure was not, since fsQCA has not yet reached it, a limitation of the audit, not evidence it lacks the same problem: an umbrella review (Bryer et al. 2025) documents comparable measurement heterogeneity there.

5. Conclusions

No published fsQCA calibration method weights by source reliability, and a 34-study audit finds applied fsQCA studies report no reliability statistic for calibration either; I introduced Extraction-Confidence-Weighted Calibration to close that gap, a simple, no-new-data addition to standard fsQCA calibration. A controlled simulation shows a bounded benefit: it reduces calibration error near the crossover but increases it for confidently-extreme cases, a boundary that behaves consistently with real data too, since gating Ragin's benchmark reverses the uniform correction's aggregate shifts by leaving unmoved exactly the confidently-extreme countries the boundary condition says a correction should leave alone. Four real, previously published datasets validate this directly, with gating variously reversing or attenuating each dataset's modest, real shifts; real data offer no ground truth to prefer one version over the other, so both are reported throughout (Sections 3.2–3.5). A further, independent check (Section 3.6) confirms that the configurational vulnerability motivating ECWC is real, not hypothetical: minimized Boolean solutions on Ragin's

own original calibration change in anywhere from roughly a third to over ninety percent of replicates, depending on perturbation level and scheme, well within an ordinary range of calibration disagreement, though whether ECWC's own correction would stabilize it remains untested.

Beyond these validations, technology-adoption and sustainability research routinely combine unequally reliable data, and the same gap already appears in public health's own fsQCA literature; nature exposure is not yet reached by fsQCA, but an independent umbrella review already documents comparable measurement heterogeneity there, so the same logic likely applies once it is. Practically, I recommend the adaptively gated version wherever the added derivation step is feasible, grounded in the simulation's own known ground truth: applied uniformly, mainly for continuity, the correction is on average worse than no correction at all; applied with the gate, it tells a researcher in advance whether the correction is warranted.

Declarations

Data availability

No new data were collected for this study. Three items will be deposited in a public repository on GitHub, archived to Zenodo, prior to publication: the simulation code and the code used to reproduce the four real-data validation analyses (drawing on publicly available data from (Ragin 2008), as reproduced in (Mendel and Korjani 2015); from Chen and Ye's (2023) publicly deposited smart-clothing purchase-intention survey; from Zheng et al.'s (2026) publicly deposited mobile fitness-app continuance-intention survey; and from the ASHRAE Global Thermal Comfort Database II (Földváry Ličina et al. 2018)), along with the code for the Boolean-minimization ground-truth validation and stability test reported in Section 3.6 and Online Resource 1 (drawing directly on Ragin's own published calibration, Ragin 2008, Ch. 5, Table 5.2); the Ragin benchmark's calibrated values, reliability weights, and derived shift tables underlying Section 3.4; and the Monte Carlo boundary-sweep outputs and further robustness-check data underlying Sections 3.1, 3.4, 4.1, and 4.5.

Competing interests

The author declares no competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

This list also includes works cited only in the accompanying Supplementary Information (Online Resource 1), which carries no reference list of its own.

Abbasi, G.A., Sandran, T., Ganesan, Y., Iranmanesh, M.: Go cashless! Determinants of continuance intention to use E-wallet apps: a hybrid approach using PLS-SEM and fsQCA. *Technol. Soc.* 68, 101937 (2022). https://doi.org/10.1016/j.techsoc.2022.101937

Ahmadov, T., Foli, S., Durst, S., Gerstlberger, W.: The transition to a circular economy: different paths for international and non-international micro-manufacturing firms. *Discov. Sustain.* 5, 178 (2024). https://doi.org/10.1007/s43621-024-00367-3

ANSI/ASHRAE: *Standard 55-2023, Thermal Environmental Conditions for Human Occupancy*. ASHRAE, Atlanta, GA (2023).

Batista Hernandez, N., Leyva Vazquez, M.Y., Abarca Cruz, A.M., Alvarado Ajila, L.A.: Measuring social emergencies through fuzzy qualitative comparative analysis: a calibration framework using normalized general continuous linguistic variables. *Front. Sociol.* 11, 1824793 (2026). https://doi.org/10.3389/fsoc.2026.1824793

Bryer, B., Odebeatu, C.C., Lee, W.R., Vitangcol, K., Gallegos-Rejas, V., Osborne, N.J., Williams, G., Darssan, D.: Greenspace exposure and associated health outcomes: a systematic review of reviews. *F1000Res* 13, 491 (2025). https://doi.org/10.12688/f1000research.148878.2

Burghgraeve, E., De Neve, J., Rosseel, Y.: Estimating structural equation models using James–Stein type shrinkage estimators. *Psychometrika* 86, 96–130 (2021). https://doi.org/10.1007/s11336-021-09749-2

Chen, J., Gu, J., Kwon, S.: Empirical Bayes shrinkage (mostly) does not correct the measurement error in regression. arXiv preprint, arXiv:2503.19095 (2025). https://doi.org/10.48550/arXiv.2503.19095

Chen, S., Ye, J.: Understanding consumers' intentions to purchase smart clothing using PLS-SEM and fsQCA. *PLoS One* 18, e0291870 (2023). https://doi.org/10.1371/journal.pone.0291870

Cheung, T., Schiavon, S., Parkinson, T., Li, P., Brager, G.: Analysis of the accuracy on PMV-PPD model using the ASHRAE Global Thermal Comfort Database II. *Build. Environ.* 153, 205–217 (2019). https://doi.org/10.1016/j.buildenv.2019.01.055

Cooper, B., Glaesser, J.: Exploring the robustness of set-theoretic findings from a large N fsQCA: an illustration from the sociology of education. *International Journal of Social Research Methodology* 19, 445–459 (2016). https://doi.org/10.1080/13645579.2015.1033799

Cronbach, L.J., Meehl, P.E.: Construct validity in psychological tests. *Psychol. Bull.* 52, 281–302 (1955). https://doi.org/10.1037/h0040957

DerSimonian, R., Laird, N.: Meta-analysis in clinical trials. *Controlled Clinical Trials* 7, 177–188 (1986). https://doi.org/10.1016/0197-2456(86)90046-2

Duşa, A.: *QCA with R: A Comprehensive Resource*. Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-319-75668-4

Efron, B., Hastie, T.: *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*, Chapter 7. Cambridge University Press, Cambridge (2016). https://doi.org/10.1017/CBO9781316576533

Földváry Ličina, V., Cheung, T., Zhang, H., de Dear, R., Parkinson, T., Arens, E., et al.: Development of the ASHRAE Global Thermal Comfort Database II. *Build. Environ.* 142, 502–512 (2018). https://doi.org/10.1016/j.buildenv.2018.06.022

Fontanari, J.F., Santos, M.: The same problem by different names: unifying regression dilution and regression to the mean. *Mathematics* 14, 2052 (2026a). https://doi.org/10.3390/math14122052

Fontanari, J.F., Santos, M.: When correcting for regression to the mean is worse than no correction at all. *Am. Nat.* 208, 314–331 (2026b). https://doi.org/10.1086/741601

He, Y., Cheng, C., Wang, L.: Unmasking the effects of e-leadership on virtual team effectiveness by an integrated fsQCA and NCA method. *PLoS One* 20, e0331500 (2025). https://doi.org/10.1371/journal.pone.0331500

Hug, S.: Qualitative comparative analysis: how inductive use and measurement error lead to problematic inference. *Polit. Anal.* 21, 252–265 (2013). https://doi.org/10.1093/pan/mps061

Kelley, T.L.: *Interpretation of Educational Measurements*. World Book Company, Yonkers-on-Hudson (1927).

Korjani, M.M., Mendel, J.M.: Interval type-2 fuzzy set qualitative comparative analysis (IT2-fsQCA). In: *Proceedings of the 2014 IEEE Conference on Norbert Wiener in the 21st Century (21CW)*, Boston, 24–26 June 2014 (2014). https://doi.org/10.1109/NORBERT.2014.6893903

Krogslund, C., Choi, D.D., Poertner, M.: Fuzzy sets on shaky ground: parameter sensitivity and confirmation bias in fsQCA. *Polit. Anal.* 23, 21–41 (2015). https://doi.org/10.1093/pan/mpu016

Kumar, S., Sahoo, S., Lim, W.M., Kraus, S., Bamel, U.: Fuzzy-set qualitative comparative analysis (fsQCA) in business and management research: a contemporary overview. *Technol. Forecast. Soc. Change* 178, 121599 (2022). https://doi.org/10.1016/j.techfore.2022.121599

Legewie, N.: Anchored calibration: from qualitative data to fuzzy sets. *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research* 18, 3 (2017). https://doi.org/10.17169/fqs-18.3.2790

Li, C.Z., Ling, X., Jiang, M., Xie, P.: Investigating environmental efficiency upgrading path of construction waste based on configuration analysis. *Sustainability* 16, 1784 (2024b). https://doi.org/10.3390/su16051784

Li, S., Zhu, J., Chen, P., Li, Z., Liao, C., Feng, W.: Exploring high-performance green innovation in China's logistics companies: a TOE framework based on fsQCA. *Front. Environ. Sci.* 12, 1511269 (2025a). https://doi.org/10.3389/fenvs.2024.1511269

Li, X., Yang, C., Liu, L., Ding, Y., Xue, J., He, J., Wu, H., Liu, L.: Configurational paths to turnover intention among primary public health workers in Liaoning Province, China: a fuzzy-set qualitative comparative analysis. *BMC Public Health* 24, 369 (2024a). https://doi.org/10.1186/s12889-024-17881-8

Li, X., Yao, Y., Shen, Q.: From dilemma to breakthrough: fuzzy-set qualitative comparative analysis of green development behavior pathways in construction materials enterprises. *Humanit. Soc. Sci. Commun.* 12, 198 (2025b). https://doi.org/10.1057/s41599-025-04493-9

Liu, R., Wei, X.: What drives green technological innovation effectiveness? A configurational analysis. *Systems* 14, 122 (2026). https://doi.org/10.3390/systems14020122

Lu, L., Ren, Y., Dong, R., Li, C., Mao, S.: Green technology upgrade path for food supply chain enterprises: insights from grounded theory and fuzzy-set qualitative comparative analysis (fsQCA). *Front. Sustain. Food Syst.* 9, 1550511 (2025). https://doi.org/10.3389/fsufs.2025.1550511

Maassen, G.H.: Kelley's formula as a basis for the assessment of reliable change. *Psychometrika* 65, 187–197 (2000). https://doi.org/10.1007/bf02294373

Maggetti, M., Levi-Faur, D.: Dealing with errors in QCA. *Political Research Quarterly* 66, 198–204 (2013).

Mendel, J.M., Korjani, M.M.: *A New Methodology for Calibrating Fuzzy Sets in fsQCA Using Level 2 and Interval Type-2 Fuzzy Sets*. USC-SIPI Report #431, University of Southern California, Los Angeles (2015).

Mendel, J.M., Korjani, M.M.: A new method for calibrating the fuzzy sets used in fsQCA. *Inf. Sci.* 468, 155–171 (2018). https://doi.org/10.1016/j.ins.2018.07.050

Miao, Z., Zhao, G.: Configurational paths to the green transformation of Chinese manufacturing enterprises: a TOE framework based on the fsQCA and NCA approaches. *Sci. Rep.* 13, 19181 (2023). https://doi.org/10.1038/s41598-023-46454-9

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

Olsen, W.: Calibration of fuzzy sets, calibration of measurement: a realist synthesis. *Systematic Mixed-Methods Research for Social Scientists*, pp. 157–174. Palgrave Macmillan, Cham (2022). https://doi.org/10.1007/978-3-030-93148-3_7

Rafe, A., Das, S.: Socio-conformal calibration in complex survey data: marginal validity is not enough for subgroup reliability. arXiv preprint, arXiv:2605.05562 (2026). https://doi.org/10.48550/arXiv.2605.05562

Ragin, C.C.: *Redesigning Social Inquiry: Fuzzy Sets and Beyond*, Chapter 5. University of Chicago Press, Chicago (2008). https://doi.org/10.7208/chicago/9780226702797.001.0001

Shafaghat, A., Hu, D., Keyvanfar, A.: A fuzzy-set qualitative comparative analysis (fsQCA) of dynamic capabilities and microfoundations for hyperscale data center project delivery. *Buildings* 16, 2284 (2026). https://doi.org/10.3390/buildings16112284

Skaaning, S.-E.: Assessing the robustness of crisp-set and fuzzy-set QCA results. *Sociol. Methods Res.* 40, 391–408 (2011). https://doi.org/10.1177/0049124111404818

Toyoda, Y.: ThSQCA: threshold-sweep qualitative comparative analysis in R. arXiv preprint, arXiv:2601.11229 (2026). https://doi.org/10.48550/arXiv.2601.11229

Wang, G., Luo, L.: Fuzzy-set qualitative comparative analysis of influencing factors on family doctor service performance during major public health emergencies. *Front. Public Health* 13, 1565499 (2025). https://doi.org/10.3389/fpubh.2025.1565499

Wang, X., Zhu, L., Tang, Y., Deng, H., Wang, H.: Multiple paths to green building popularization under the TOE framework: a qualitative comparative analysis of fuzzy sets based on 26 Chinese cities. *Sustainability* 16, 9360 (2024). https://doi.org/10.3390/su16219360

WCED: *Our Common Future*. Oxford University Press, Oxford (1987).

Wei, J., Liu, F.: Policy pathways to digitalization: a qualitative comparative analysis of China's construction industry. *Buildings* 16, 1772 (2026). https://doi.org/10.3390/buildings16091772

Willink, R.: Shrinkage estimation to minimize error in measurement estimates and consensus values. *Metrology* 6, 26 (2026). https://doi.org/10.3390/metrology6020026

Zheng, Y., Li, K., Wu, Q.: Determinants of mobile fitness app continuance intention using PLS-SEM, NCA and fsQCA. *Sci. Rep.* 16, 23972 (2026). https://doi.org/10.1038/s41598-026-55404-0

Zhou, Y., Chen, L., Zhou, X., Hu, S.: Key factors and path selection for enterprise digital transformation: configuration analysis based on fsQCA. *Front. Sustain.* 5, 1366129 (2024). https://doi.org/10.3389/frsus.2024.1366129

Supplementary Files

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

- ECWCQQSupplementary.docx