

RESEARCH ARTICLE

Equivalence Concordance Rate (ECR): A Framework for Comparing and Calibrating Equivalence Acceptance Criteria

Padraic Nagle | Huruy Asfha | Qing Meng | Emmanuel Atindama

¹Global Statistics, MS&T, Bristol Myers Squibb

Correspondence

Padraic Nagle

Email: statistics@bms.com

Abstract

Assessing a manufacturing process change often involves demonstrating equivalence between the pre- and post-change processes. The two one-sided tests (TOST) procedure is a common way to do so, with the verdict hanging on a single number, the Equivalence Acceptance Criterion (EAC): set it too wide and a shifted process is wrongly accepted, set it too narrow and an acceptable process is wrongly rejected. Many methods for setting the margin coexist, yet they are seldom ranked by the *coherence* of the decisions they induce. We introduce the **Equivalence Concordance Rate (ECR) framework**, which classifies the joint outcome of a difference test and a TOST via Monte Carlo simulation across process shifts and summarizes this outcome as three rates—the concordant (ECR), paradox (PR), and gray-zone (GZR) rates—and a single scalar, the integrated ECR (iECR), for ranking and calibrating *any* margin. The diagnostic sweeps hypothetical shifts around an estimated process, hence suitable for a single comparability dataset. We illustrate this with a case study of four common EAC methods across a constrained factorial design (10,368 valid scenarios) and derive a closed-form off-center formulation of the capability-informed (Ppk) margin. Additionally, we quantify the drivers of margin performance: sample size dominates, followed by the method $\times$ standard-deviation-ratio interaction.

Finally, we provide an application of the framework to real manufacturing data in the appendix.

KEYWORDS

equivalence, concordance, paradox, diagnostic, framework, process comparability, operating characteristics, Monte Carlo Simulation

1 | INTRODUCTION

Pharmaceutical manufacturing process changes are commonly assessed for equivalence between the pre- and post-change processes via the two one-sided tests (TOST) procedure with a pre-specified margin (EAC)^{1–6}. Too wide a margin declares equivalence for significant shifts; too narrow a margin fails to demonstrate it for processes that are comparable. Guidance such as ICH Q5E specifies no single statistical method, and comparability may rest on analytical characterization alone, with nonclinical or clinical studies warranted where analytical evidence cannot establish quality, safety, and efficacy equivalence⁵.

Existing operating-characteristic (OC) analyses rank equivalence procedures and margins by their error rates^{7,8}, and regulatory guidance and related literature recommend such risk-based justification^{9–11}. These score the equivalence test

against an externally fixed standard of which true shifts are acceptable. They do not ask whether a separate difference test would *agree* with that verdict, nor do they rank the *construction methods* that generate margins^{12–16}.

We employ the four-zone classification of joint difference/equivalence outcomes originated in three-sided testing¹⁷ and related schemes^{18–20}. Conditional equivalence testing reaches the same conclusions by testing equivalence only after a non-significant difference test²¹, which forecloses the paradox rather than measuring it. Other works replace the margin decision instead: second-generation *p*-values^{22–24}, data-dependent post-hoc margins²⁵, and Bayesian HDI–ROPE comparisons^{26–29}. However, they do not score margins by the coherence of their verdicts. The closest study⁸ does score margins, ranking candidate EAC *values* by the equivalence test's false-acceptance and false-rejection rates. We instead rank the construction *methods*, by their agreement with a difference test.

We propose the **Equivalence Concordance Rate (ECR) framework** to fill this gap; relabeling the four zones as concordant-equivalent, concordant-different, gray zone, and paradox^{17,18,20}. We compute their *frequencies* via Monte Carlo as a function of the true shift δ , giving three diagnostic curves—concordant (ECR), paradox (PR), and gray-zone (GZR) rates—and a summary scalar, the integrated ECR (iECR).

Contributions.

Our primary contribution is this framework: a method-agnostic diagnostic that scores how well *any* margin supports coherent decisions and calibrates the tuning parameters to maximize iECR, with paradox and gray-zone trade-offs. By sweeping hypothetical shifts around an estimated process, a single comparability dataset is sufficient to benchmark any incumbent margin.[†] Three further contributions follow: (i) a *case study* of four common EAC methods at prespecified illustrative settings, with a proposed decision framework; (ii) a closed-form, off-center formulation of the capability-informed margin, parameterized by a capability retention fraction $\rho \in (0, 1)$ ($Ppk_{target} = \rho Ppk_{pre}$); and (iii) a factorial ANOVA that identifies margin performance drivers.[‡]

2 | THE EQUIVALENCE CONCORDANCE RATE (ECR) FRAMEWORK

A well-calibrated margin maximizes agreement between the two verdicts; the rate of disagreement—contradiction or inconclusiveness—measures its calibration. The ECR framework turns this intuition into a measurement, returning for any candidate margin Δ a quantitative diagnostic of its calibration quality and trade-offs. We first fix notation and define the metrics.

2.1 | Notation and Assumptions

Consider a manufacturing process before and after a change, with

$$X_i \sim \mathcal{N}(\mu_{pre}, \sigma_{pre}^2), \quad i = 1, \dots, n_{pre}, \quad (1)$$

$$Y_j \sim \mathcal{N}(\mu_{pre} + \delta_{true}, \sigma_{post}^2), \quad j = 1, \dots, n_{post}, \quad (2)$$

where μ_{pre} and σ_{pre} are the pre-change mean and standard deviation and δ_{true} is the *true* (unknown) mean shift caused by the change. The product attribute is subject to two-sided specification limits $[LSL, USL]$ [§].

[†] See application to real data in Appendix H

[‡] See Appendix B.

[§] Specification limits may not apply to every process-change attribute. They are used only by the Ppk-Informed margin (Method D).

Two hypothesis tests are applied to each two-sample dataset: A *difference test* (D-test), a two-sided *t*-test at significance level α_d to check significant difference:

$$H_0: \mu_{post} - \mu_{pre} = 0 \quad \text{vs} \quad H_a: \mu_{post} - \mu_{pre} \neq 0. \quad (3)$$

An *equivalence test* (E-test), the TOST procedure at level α_e , to test equivalence within the margin $\pm\Delta$ [¶]:

$$H_0: |\mu_{post} - \mu_{pre}| \geq \Delta \quad \text{vs} \quad H_a: |\mu_{post} - \mu_{pre}| < \Delta, \quad (4)$$

where $\Delta > 0$ is the equivalence acceptance criterion (EAC).

Both tests are computed from the estimated mean difference $\hat{\delta} = \bar{Y} - \bar{X}$ and the pooled standard error

$$\begin{aligned} se &= s_p \sqrt{1/n_{pre} + 1/n_{post}}, \\ s_p &= \sqrt{\frac{(n_{pre} - 1)s_{pre}^2 + (n_{post} - 1)s_{post}^2}{n_{pre} + n_{post} - 2}}, \end{aligned} \quad (5)$$

with $df = n_{pre} + n_{post} - 2$. Observations are assumed independent and normally distributed with equal population variances ($\sigma_{post}/\sigma_{pre} \approx 1$), and $\alpha_d = \alpha_e = 0.05$ unless stated otherwise.

2.2 | Decision Matrix and Diagnostic Rates

The joint outcome of the two tests on the same data falls into one of four mutually exclusive zones (Table 1). This cross-classification appears in three-sided testing¹⁷ and the combined significance/equivalence schemes of^{18,20}. Our contribution is to turn it into an operating characteristic of the joint procedure, estimating the four zone probabilities by Monte Carlo at each shift δ to obtain the ECR, GZR, and PR curves:

	E-test pass (equivalence concluded)	E-test fail (equivalence not concluded)
D-test fail (no significant difference)	Z ₁ ✓ Concordant-equivalent	Z ₂ ? Gray zone
D-test pass (significant difference)	Z ₃ ! Paradox	Z ₄ ✓ Concordant-different

TABLE 1 Decision matrix notation. Z₁–Z₄ denote zone counts from N Monte Carlo replicates.

From this classification we define three diagnostics:

$$\begin{aligned} ECR &= (Z_1 + Z_4)/N && \text{(concordant decisions)} \\ PR &= Z_3/N && \text{(paradox: verdicts conflict)} \\ GZR &= Z_2/N && \text{(gray zone: inconclusive)} \end{aligned}$$

[¶] In plain terms: is the shift small enough to be practically equivalent?

By construction $ECR + PR + GZR = 1$ at every shift. A high ECR across the shift range indicates a well-calibrated margin; elevated PR, a margin wide relative to detectability; elevated GZR, a margin that is too narrow or underpowered for a decisive outcome.

These rates are *not* the false-acceptance/false-rejection operating characteristics of the equivalence test alone^{7,8}, which judge a single TOST decision against the ground truth. Because the difference test has a point null ($H_0: \delta = 0$) and flags *any* detectable shift, a paradox can arise even when TOST *correctly* declares equivalence for a shift inside the margin: PR measures *disagreement*, not a false-acceptance rate. The false positive/negative OC analyses in⁸ must fix externally which true differences count as acceptable. ECR instead references statistical *detectability*. Where a defensible relevance threshold exists it should be used directly; otherwise the difference test is a stand-in whose validity weakens at large n [#]. Hence these two are complementary.

The two disagreement modes are not defects but opposite directions of decision risk: a paradox is the *consumer's-risk* direction, a gray zone the *producer's-risk* direction (an acceptable process left unconfirmed, a cost to the manufacturer³⁰). Reporting them separately lets a practitioner weigh both rather than collapsing them into one verdict. Neither is intrinsically an error^{18,26}: an elevated PR indicates a margin wide relative to detectability—not necessarily a practically harmful change, which follows only insofar as the point-null test is a defensible relevance standard.^{||}

2.3 | The ECR Curve: $ECR(\delta/\sigma_{pre})$

The ECR curve traces the proportion of concordant decisions against the standardized true shift:

- *Small shifts* ($\delta \ll \Delta$): no difference is detected and, when TOST is adequately powered, equivalence is concluded; the outcome is predominantly Z_1 and ECR approaches $1 - \alpha_d$.
- *Large shifts* ($\delta \gg \Delta$): a difference is detected and equivalence is not concluded; again the tests agree and $ECR \approx 1$.
- *Intermediate shifts*: $\hat{\delta}$ falls between the two thresholds, the tests frequently disagree, and the curve dips—near, but not exactly at, $\delta = \Delta$, since the disagreement window is bounded by t_d se and $\Delta - t_e$ se, not by Δ itself.

The dip is the diagnostic signature of the method.

The integrated ECR (iECR).

For ranking, we define the *integrated ECR* as the average height of the ECR curve over the standardized shift range. With $u = \delta/\sigma_{pre}$ and $u_{max} = \delta_{max}/\sigma_{pre}$,

$$iECR = \frac{1}{u_{max}} \int_0^{u_{max}} ECR(u) du. \quad (6)$$

A higher iECR indicates better calibration, so methods can be ranked by this one number. The shift limit changes the absolute level and can reorder the lower-ranked methods, but not which method ranks highest.^{**}

3 | EAC METHODS: A CASE STUDY

To demonstrate the framework end-to-end, we apply it to four commonly used methods for computing the EAC. Each encodes a different philosophy of a tolerable shift: process variability (A), the process mean (B), an assumed effect size with a finite-sample variability correction (C), and process capability (D).

Method A: σ -Scaled.

$$EAC_A = \lambda \cdot \sigma_{pre} \quad (7)$$

The margin scales linearly with pre-change variability by a pre-specified λ . The clearest published precedent for a specific value is regulatory: the FDA's draft Tier-1 criterion for analytical similarity set the margin at $1.5 \sigma_R$, where σ_R is the reference-product standard deviation^{31,††}. This calibration was for biosimilar analytical similarity, not process comparability. No application-agnostic basis for choosing λ has been established. The ECR framework (Section 2) provides one.

Method B: Percent-of-Mean.

$$EAC_B = \lambda_{\%} \cdot |\mu_{pre}| \quad (8)$$

This margin is a fixed percentage of the process mean. Mean-referenced criteria are well established in pharmaceutical practice: bioequivalence requires the 90% confidence interval for the geometric mean ratio to lie within 80.00–125.00%³², though there the limits apply to a ratio on the log scale. In σ units the margin is $\lambda_{\%} \mu_{pre}/\sigma_{pre}$, so it tracks the process mean rather than variability. It is therefore too wide at large μ_{pre}/σ_{pre} and too tight at small values.^{‡‡}

^{**} Appendix D: Method A is highest at every δ_{max} ; C and D come within under 0.002 of each other and cross, too close to separate.

^{††} That 2017 draft was withdrawn in 2018.

^{‡‡} This swing is why it sits at the lower end of the grid comparison in Section 5.2.

[#] See Appendix C

^{||} Section 6 discusses a practical-relevance gate that restores that reading at large n .

Method C: Effect-Size-Based.

$$\text{EAC}_C = ES \cdot \hat{\sigma}_{upper}, \quad ES = \frac{\delta_{assumed}}{\sigma_{pre}},$$

$$\hat{\sigma}_{upper} = \sqrt{\frac{(n_{pre} - 1)s_{pre}^2}{\chi_{\alpha_{\chi^2}, n_{pre}-1}^2}} \quad (9)$$

Here $\delta_{assumed}$ is the analyst's *a priori* assumed shift, a tuning parameter distinct from the true shift δ_{true} . The bound $\hat{\sigma}_{upper}$ replaces σ_{pre} with its $(1 - \alpha_{\chi^2})$ -level upper confidence bound, which accounts for sampling variability¹⁶, Sec. 9.4.1. Related finite-sample corrections appear in³³. A smaller α_{χ^2} inflates $\hat{\sigma}_{upper}$ and widens the margin. The inflation factor approaches 1 as n_{pre} grows.³³ The margin is a product of a judgment (ES) and an estimate (s_{pre}), so an error in either scales the whole margin.

Remark 1. Relationship to Method A (σ -Scaled). Set $c = \sqrt{\frac{n_{pre}-1}{\chi_{\alpha_{\chi^2}, n_{pre}-1}^2}} \geq 1$ and evaluate $\hat{\sigma}_{upper}$ at $s_{pre} = \sigma_{pre}$:

$$\text{EAC}_C = \underbrace{ES}_{\lambda} \cdot \underbrace{\sigma_{pre} \cdot c}_{\hat{\sigma}_{upper}} \xrightarrow{n_{pre} \rightarrow \infty} \underbrace{ES}_{\lambda} \cdot \sigma_{pre} = \text{EAC}_A.$$

With $ES = \lambda$, Method C is Method A inflated by c : it widens the margin to absorb sampling variability from finite data, where Method A takes σ_{pre} at face value. That is their only structural difference. Outside the simulation σ_{pre} is itself estimated, so $\text{EAC}_C = ES \cdot s_{pre} \cdot c$ and the identity holds as a probability limit rather than an exact equality.

Method D: Ppk-Informed.

Unlike the preceding methods, Method D anchors the margin to process capability via the performance index Ppk ³⁰. The index is introduced by example in¹⁶, Sec. 9.4.1. The closed-form expression, off-center generalization, and derivation below are contributions of this paper. Given a target capability Ppk_{target} to be maintained after the change, the largest absorbable mean shift defines the margin. We derive that margin from the definition of Ppk , then generalize it to the off-center case (Corollary 1).

Definition 1 (Process Performance Index). For a process with mean μ and standard deviation σ within limits $[LSL, USL]$,

$$Ppk = \min\left(\frac{USL - \mu}{3\sigma}, \frac{\mu - LSL}{3\sigma}\right). \quad (10)$$

Proposition 1 (Ppk-Informed EAC (Centered Case)). Consider a centered pre-change process with mean μ_{pre} , standard deviation σ_{pre} , and specification limits $[LSL, USL]$. Write $S = (USL - LSL)/2$ for the specification half-width. Suppose the

process undergoes a mean shift δ with unchanged variance, $\sigma_{post} = \sigma_{pre}$.^{¶¶} Then the largest tolerable shift for which $Ppk_{post} \geq Ppk_{target}$, which defines the margin, is

$$\text{EAC}_D = \max(S - 3 \cdot Ppk_{target} \cdot \sigma_{pre}, 0). \quad (11)$$

Corollary 1 (Off-Center Correction). When $\mu_{pre} \neq \frac{USL+LSL}{2}$, the process sits closer to one specification limit than the other. The relevant distance is then not S but d_{pre} , the distance from μ_{pre} to the nearest specification limit,

$$d_{pre} = \min(USL - \mu_{pre}, \mu_{pre} - LSL) \leq S.$$

For $Ppk_{target} < Ppk_{pre}$, the Ppk-Informed EAC that guarantees $Ppk_{post} \geq Ppk_{target}$ regardless of shift direction is then

$$\text{EAC}_D = \max(d_{pre} - 3 Ppk_{target} \sigma_{pre}, 0). \quad (12)$$

This robust form reduces to (11) when the process is centered ($d_{pre} = S$). It is the operative definition of EAC_D in the remainder of the paper. The clip at zero carries no such guarantee. Once $Ppk_{target} \geq Ppk_{pre}$ the target is unmet even at $\delta = 0$, so $\text{EAC}_D = 0$ records infeasibility rather than a usable margin (Remark 2). The capability guarantee is conditional on an unchanged spread. If $\sigma_{post} > \sigma_{pre}$, then in the worst case $Ppk_{post} = (d_{pre} - |\delta|)/(3\sigma_{post})$, so $Ppk_{post} \geq Ppk_{target}$ can fail even for a shift within the margin. Section 5.3 probes this by varying $\sigma_{post}/\sigma_{pre}$.

The proofs of Proposition 1 and Corollary 1 are shown in Appendix A.

Remark 2. Feasibility Boundary and Capability Retention. The robust margin (12) decreases linearly in Ppk_{target} and vanishes once

$$Ppk_{target} \geq \frac{d_{pre}}{3\sigma_{pre}} = Ppk_{pre}, \quad (13)$$

beyond which no usable margin exists (Figure 1). At equality only $\delta = 0$ meets the target, and a zero-width margin leaves the TOST alternative empty, so no shift can be declared equivalent. Above it the target fails even at zero shift. Anchoring directly to current capability ($Ppk_{target} = Ppk_{pre}$) therefore collapses the margin to zero. To retain a usable margin we scale by a *capability retention fraction* $\rho \in (0, 1)$, setting $Ppk_{target} = \rho Ppk_{pre}$:

$$\text{EAC}_D = \max(d_{pre} - 3\rho Ppk_{pre} \sigma_{pre}, 0) = d_{pre} (1 - \rho). \quad (14)$$

All experiments adopt $\rho = 0.8$ unless stated otherwise. A fixed quality or business target such as $Ppk_{target} = 1.33$ may be used instead when a specific capability requirement applies.

^{§§} At $\alpha_{\chi^2} = 0.05$ it is $\approx 1.6 \times$ at $n_{pre} = 10$ and $\approx 1.3 \times$ at $n_{pre} = 30$.

^{¶¶} Mirroring the equal-variance assumption of the two tests, this lets the post-change capability be written in terms of the known σ_{pre} ; Section 5 relaxes it to probe robustness.

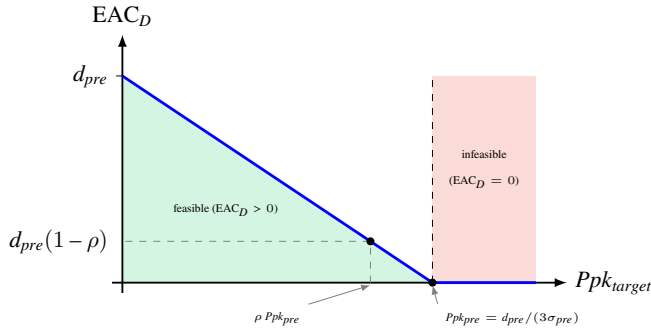


FIGURE 1 Feasibility boundary for the Ppk-Informed margin: EAC_D against the target capability Ppk_{target} . The boundary sits at the pre-change capability $Ppk_{pre} = d_{pre}/(3\sigma_{pre})$; the shaded region is infeasible. The retention fraction ρ selects an operating point inside the feasible region, shown for $\rho = 0.8$.

4 | EXPERIMENTAL SETUP

4.1 | Monte Carlo Simulation Engine

Consider a pre-change population with parameters μ_{pre} and σ_{pre} , specification limits $[LSL, USL]$, and a pre-change sample size n_{pre} . Each method's margin Δ_M is computed from these pre-change parameters, once per scenario. No margin uses a simulated sample. For Method C the χ^2 bound is evaluated at $s_{pre} = \sigma_{pre}$ (Remark 1), so $\hat{\sigma}_{upper} = c \sigma_{pre}$ is deterministic.

With Δ_M fixed, the engine repeats the following N_{sim} times at each shift δ_{true} .

1. Draw n_{pre} observations from $\mathcal{N}(\mu_{pre}, \sigma_{pre}^2)$.
2. Draw n_{post} observations from $\mathcal{N}(\mu_{pre} + \delta_{true}, \sigma_{post}^2)$.
3. Compute $\hat{\delta}$ and se from the two samples.
4. Apply the D-test at α_d and TOST at α_e against Δ_M .
5. Classify the outcome into one of Z_1 – Z_4 (Table 1).

The zone proportions over the N_{sim} replicates estimate ECR , GZR , and PR .

We resample per replicate, thus $\hat{\delta}$ and se vary, while Δ_M is fixed. The margin, and hence the analysis is therefore *dependent* on the pre-change parameters, but *independent* of any specific sample. One consequence is specific to Method C: its χ^2 inflation hedges against underestimating σ_{pre} , so a known σ_{pre} leaves that hedge unexercised and part of its paradox rate is an artifact of the design. Appendix F repeats the full grid so that margins are dependent on pre-change samples.

All methods use the *same* simulated datasets, with common random numbers and only Δ_M differing. This isolates

margin effects from Monte Carlo noise. The pooled t -test is used throughout ($df = n_{pre} + n_{post} - 2$). It is retained even when $\sigma_{post}/\sigma_{pre} \neq 1$, to test the diagnostics under a violated equal-variance assumption (Section 5).

4.2 | Factorial Design

Off-Center Process Mean.

In practice, a manufacturing process is never perfectly centered between its specification limits. Let $m = \frac{USL+LSL}{2}$ be the specification midpoint. We parameterize the process mean as

$$\mu_{pre} = m + \kappa \cdot S, \quad (15)$$

where κ is the *off-center fraction*. The nearer-limit distance is then $d_{pre} = S(1-|\kappa|)$ (Remark 2). The grid samples only $\kappa > 0$. Thus, a wider percent-of-mean margin since it is sensitive to the *sign* of κ .

Sample-Size Constraints.

Pre-change sizes n_{pre} range from 4 to 30, reflecting typical pharmaceutical batch sizes. Post-change sizes are $n_{post} = \lfloor r \cdot n_{pre} \rfloor$, rounded to the nearest integer, for $r \in \{1/3, 1/2, 1, 2\}$ subject to $4 \leq n_{post} \leq 30$. This captures the asymmetry often between historical dataset and post-change batch. Not every $n_{pre} \times r$ pair is feasible:

- $r = 2$ only for $n_{pre} \leq 12$.
- $r = 1/3$ at $n_{pre} \leq 8$ and $r = 1/2$ at $n_{pre} = 4$.
- Twenty-four of the 32 pairs survive, making the design a *constrained factorial grid*.

The feasible set of r varies with n_{pre} , so marginal n_{pre} curves average over slightly different sets (Appendix B provides the contribution of the sample size ratio).

Factorial Grid.

The EAC formulae use the pre-change parameters only through ratios, and the t -statistics are scale-invariant. We hereby fix $\sigma_{pre} = 1$ and vary six drivers: m/σ_{pre} , S/σ_{pre} , κ , n_{pre} , n_{post}/n_{pre} , and $\sigma_{post}/\sigma_{pre}$. Which driver matters for which method is left to the results (Section 5, Appendix B). Table 2 lists the values.

The grid gives $6 \times 3 \times 8 \times 4 \times 6 \times 4 = 13,824$ nominal combinations, of which 10,368 are valid. The ranges span realistic process conditions:

- m/σ_{pre} spans 5–50.
- $S/\sigma_{pre} \in \{2.5, 5, 10\}$ corresponds to centered $Ppk \in \{0.83, 1.67, 3.33\}$. Off-center scenarios scale these by $(1 - \kappa)$, since $Ppk_{pre} = (1 - \kappa) S/(3\sigma_{pre})$.

TABLE 2 Factorial grid: dimensionless process parameters. All quantities except n_{pre} are ratios; $\sigma_{pre} = 1$ is fixed without loss of generality.

Parameter	Symbol	Grid values	Count
Midpoint-to-variability ratio	m/σ_{pre}	5, 10, 15, 25, 35, 50	6
Spec-to-variability ratio	S/σ_{pre}	2.5, 5, 10	3
Pre-change sample size	n_{pre}	4, 8, 12, 16, 20, 24, 28, 30	8
Sample-size ratio	n_{post}/n_{pre}	$\frac{1}{3}, \frac{1}{2}, 1, 2$	4
Standard-deviation ratio	$\sigma_{post}/\sigma_{pre}$	0.5, 0.75, 1.0, 1.25, 1.5, 2.0	6
Off-center fraction	κ	0.05, 0.10, 0.20, 0.35	4

- $\sigma_{post}/\sigma_{pre}$ from 0.5 to 2.0 covers both reduction and inflation of the post-change spread.
- Three of the 18 midpoint-spec width combinations ($m/\sigma_{pre} \times S/\sigma_{pre}$) pairs have $m \leq S$, hence a non-positive LSL ; they are retained for completeness.

For each combination, ECR, GZR, and PR are computed at 20 equally spaced shifts $\delta_{true}/\sigma_{pre} \in [0, 4.0]$ with $N_{sim} = 5,000$ replicates per shift.

Tuning-Parameter Adaptation.

Tuning parameters are held at fixed, prespecified defaults, so the comparison reflects one consistent as-applied configuration of each method. These are illustrative choices rather than established standards: only Method A's λ has an external anchor, Method C's ES is matched to it, and Methods B and D take round conventional values. The margins still adapt to each scenario through their dependence on σ_{pre} , μ_{pre} , n_{pre} , and Ppk_{pre} :

- Method A: $\lambda = 1.5$ (matching the FDA draft Tier-1 criterion³¹);
- Method B: $\lambda_{\%} = 0.05$ (a moderate 5% margin);
- Method C: $ES = 1.5$ (matched to Method A's λ for a like-for-like comparison), $\alpha_{\chi^2} = 0.05$;
- Method D: $Ppk_{target} = 0.8 Ppk_{pre}$ (retain $\geq 80\%$ of pre-change capability, see Remark 2).

Tuning-Parameter Optimization.

To separate *structural* performance from tuning effects, we also maximize the subset-averaged iECR over each method's primary tuning parameter. The subset fixes $n_{pre} \in \{8, 20, 30\}$, $\sigma_{post}/\sigma_{pre} = 1$, and $\kappa = 0.05$, and spans all $m/\sigma \times S/\sigma$ combinations with every feasible ratio. A coarse sweep locates each optimum. A 15-point fine grid centered on it then refines the estimate under a lighter budget ($N_{sim} = 1,000$, 8 shifts). These grid-averaged optima are a diagnostic upper bound, not

recommended settings.^{##} The prespecified-default comparison remains the primary result (Section 5.4).

4.3 | Baseline Parameterization

The baseline scenario sets:

$m/\sigma_{pre} = 50$, $S/\sigma_{pre} = 5$, $n_{pre} = n_{post} = 20$, $\sigma_{post}/\sigma_{pre} = 1.0$, and $\kappa = 0.05$. With $\sigma_{pre} = 1$ these imply

- specification limits $50 \pm 5 = [45, 55]$;
- a process mean $\mu_{pre} = m + \kappa S = 50.25$;
- a nearer-limit distance $d_{pre} = \min(USL - \mu_{pre}, \mu_{pre} - LSL) = 4.75$;
- a pre-change capability $Ppk_{pre} = (1 - \kappa) S / (3\sigma_{pre}) = 1.58$, below the centered 1.67 because $\kappa = 0.05$.

This scenario is one of the 10,368 grid combinations. It is shown on its own because the grid reports only a scalar per scenario as an average over the shifts (δ). The shape of $ECR(\delta)$ is visible only at a fixed parameterization. The large m/σ_{pre} is deliberate: it places Method B's mean-scaled margin at the wide end of its range, making that method's dependence on μ_{pre} visible. Tuning parameters are the prespecified defaults listed above, giving $Ppk_{target} = 0.8 Ppk_{pre} \approx 1.27$ for Method D. The resulting standardized margins are given in Table 4.

5 | RESULTS

Section 5.1 examines a single scenario, the baseline of Section 4.3, under two views: the two tests' decision probabilities and the ECR diagnostics. Section 5.2 then aggregates over the full grid, of which that scenario is one cell. Sections 5.3 and 5.4 follow with robustness to the design factors and each method's tuning optimum. The margins and decision probabilities are analytical³⁴; everything else is Monte Carlo. All test decisions depend only on dimensionless ratios (Section 4.2), so these results hold at any absolute scale.

5.1 | The Baseline Scenario

Test decision probabilities.

The margins order tightest to widest as $D < A < C < B$. A and C differ only by the χ^2 correction.

^{##} Run at one process's own pre-change parameters, the same sweep becomes a planning tool: it uses pre-change data only, as every EAC formula here does, and so adds no dependence on the comparability data. The value it selects must still be prespecified.

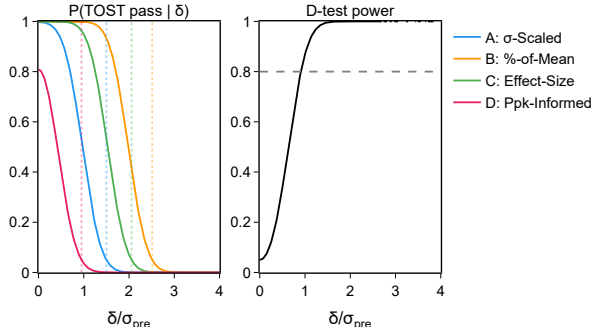


FIGURE 2 Probability that each test rejects, at baseline ($n_{pre} = n_{post} = 20$, $\sigma_{post}/\sigma_{pre} = 1$). Left: TOST pass probability vs. true shift δ ; dotted lines mark each method’s EAC. Right: D-test power (common to all methods); dashed line at 80%.

Figure 2 gives the probability of declaring equivalence (left) and of detecting difference (right). As the shift grows, equivalence becomes less likely and a difference becomes more likely. The dotted lines mark each method’s margin. Where a curve crosses its own dotted line, $\delta = \Delta_M$ and the chance of declaring equivalence is at most $\alpha_e = 5\%$. The D-test is calibrated at the other end, rejecting $\alpha_d = 5\%$ of the time at zero shift. Method D sits low even at zero shift, because its tight margin is underpowered at $n = 20$. Table 3 gives the size needed for 80% and 90% power². A tighter margin needs more data to achieve the same power. These hold at the baseline margins; scenario-specific planning should recompute power.

TABLE 3 Minimum sample size per group ($n_{pre} = n_{post}$) for target TOST power at $\delta_{true} = 0$, computed analytically from the exact joint TOST acceptance probability³⁴. Each margin is fixed at baseline value, so only sample size varies.

Method	EAC/ σ_{pre}	n for 80% power	n for 90% power
A (σ -Scaled)	1.50	9	11
B (%-of-Mean)	2.51	4	5
C (Effect-Size)	2.06	5	6
D (Ppk-Informed)	0.95	20	25

ECR diagnostics.

Before collating results across the grid, we show the diagnostics at one scenario. The baseline parameters go through the Monte Carlo engine of Section 4.1: $N_{sim} = 10,000$ replicates at each of $N_\delta = 40$ shifts over $\delta/\sigma_{pre} \in [0, 4]$. The curves above are analytic; these are simulated.

Figure 3 and Table 4 show the three diagnostics. Each ECR curve is *U-shaped*: it dips near (but not exactly at) the method’s own margin, then returns toward 1 at large shifts. A wide margin trades gray zone for paradox; a tight one does the reverse. B’s mean-scaled margin and C’s variance correction widen

TABLE 4 Baseline margins and scalar summary of the ECR curves in Figure 3. iECR is the area-normalized integral of the ECR curve over $\delta/\sigma_{pre} \in [0, 4]$, and the rest report extrema. Methods are ordered by iECR (descending).

Method	EAC/ σ_{pre}	iECR	min ECR	max GZR	max PR
D: Ppk-Informed	0.95	0.945	0.737	0.259	0.004
A: σ -Scaled	1.50	0.916	0.606	0.001	0.394
C: Effect-Size	2.06	0.777	0.169	0.000	0.831
B: %-of-Mean	2.51	0.662	0.039	0.000	0.961

theirs most. The curves overlap in places, and no method leads on all three diagnostics. The iECR order here is also not the grid’s (Section 5.2). This scenario shows shape, not rank.

5.2 | Comparison Across the Factorial Grid

This and the following two subsections use Monte Carlo across the full grid. Table 5 gives across-scenario summaries, ordered by mean iECR. Its max GZR and max PR columns are means of each scenario’s peak; EAC/ σ is the grid mean, not the baseline value in Table 4.

Mean iECR spans 0.800–0.872, with Method A highest. The other three fall within 0.022 of one another and their order is not stable: C and D differ by only 0.014 and reverse under other scenario mixes and shift limits (Appendix D). Per scenario, A has the highest iECR in 47% of the grid and the lowest in 2%; the baseline of Section 4.3 is one where it is not highest. B, C, and D are each lowest in 28–36% of scenarios, so none is reliably the weakest. C has the highest paradox rate (mean max PR = 0.512), D the highest gray-zone rate (mean max GZR = 0.596). Figure 4 shows the full distributions. Method A places most mass near 0.95 with a long lower tail, C sits slightly lower but reaches the smallest values, and D and B are broadest, centered near 0.8.

5.3 | Robustness Analysis

We vary the design factors with the largest effect sizes in the Type II analysis of variance of the factorial iECR (Appendix B).

TABLE 5 Method comparison across the constrained factorial grid (all 10,368 valid scenarios $\times$ 4 methods; $N_{sim} = 5,000$, $N_\delta = 20$ shifts over $[0, 4\sigma_{pre}]$).

Method	Mean iECR	SD	min iECR	iGZR	iPR	max GZR	max PR	EAC/ σ
A: σ -Scaled	0.872	0.107	0.397	0.092	0.036	0.220	0.170	1.50
C: Effect-Size	0.822	0.101	0.337	0.013	0.165	0.028	0.512	2.35
D: Ppk-Informed	0.808	0.112	0.381	0.172	0.020	0.596	0.083	0.96
B: %-of-Mean	0.800	0.112	0.381	0.146	0.054	0.501	0.171	1.22

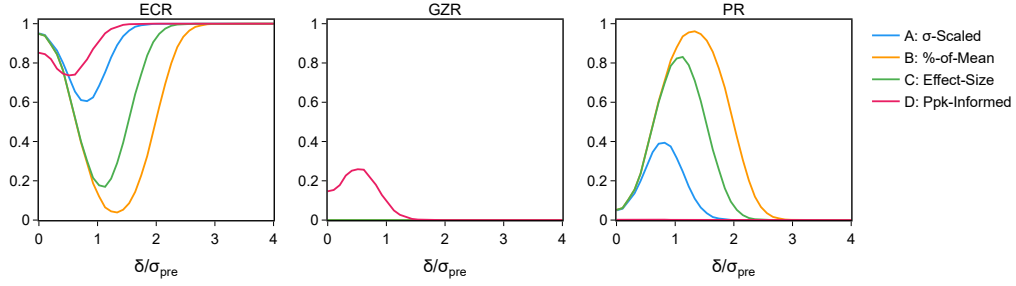


FIGURE 3 Monte Carlo ECR, GZR, and PR curves (left to right) at baseline: $m/\sigma = 50$, $S/\sigma = 5$, $n_{pre} = n_{post} = 20$, $\sigma_{post}/\sigma_{pre} = 1$, $\kappa = 0.05$; $N_{sim} = 10,000$, $N_\delta = 40$.

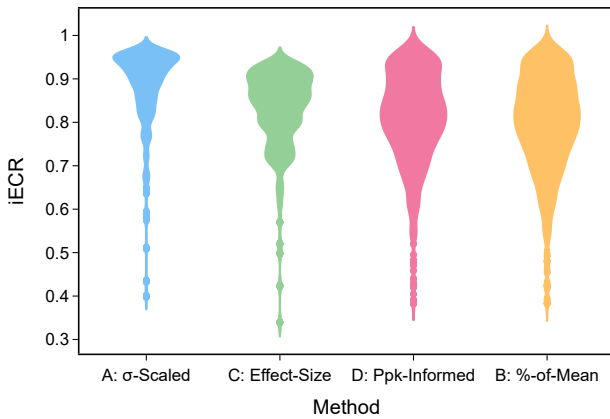


FIGURE 4 Violin plots of iECR across all 10,368 factorial scenarios (Monte Carlo, $N_{sim} = 5,000$, $N_\delta = 20$). The box shows the interquartile range, and the horizontal line marks the mean.

Sample size.

All methods improve steeply to $n_{pre} \approx 12$. Beyond that B and D continue to improve, whereas C turns over near 16 and A near 24: the overpowered point-null D-test widens the paradox window for the wider margins at large n (Appendix C). The turnover locations are approximate, as the feasible set of r varies with n_{pre} (Section 4.2).

Unequal variance.

A, B, and D decline as $\sigma_{post}/\sigma_{pre}$ grows, while C improves and overtakes A beyond ≈ 1.5 . Inflating σ_{post} enlarges se, widening the threshold separation ($t_d + t_e$) se of Section 2.2 and pushing every method toward the gray zone. That harms tight margins but corrects C, whose margin is the widest and most paradox-prone. This is the Method $\times$ standard-deviation-ratio interaction, the second-largest ANOVA effect.

Pooled versus unequal-variance tests.

Both tests pool variances throughout ($df = n_{pre} + n_{post} - 2$), which is misspecified when $\sigma_{post} \neq \sigma_{pre}$. Repeating the marginal analysis with Welch versions of both tests (Appendix E) leaves the ranking unchanged: Method C still overtakes A near 1.5. The correction lowers absolute iECR modestly under strong inflation. Method C's robustness is therefore genuine, not an artifact of pooling.

5.4 | Tuning-Parameter Optima

Figure 6 reports the sweep of each method's primary tuning parameter over the factorial subset (Section 4.2). Tuning does not make the methods converge; the peaks stay well separated. Each method's iECR-maximizing setting θ^* (marked in Figure 6) and its peak iECR are C: $ES^* = 1.00$ (0.930), A: $\lambda^* = 1.43$ (0.908), D: $\rho^* = 0.83$ (0.842), and B: $\lambda_{\%}^* = 0.039$ (0.836). Each optimum is a broad plateau rather than a sharp point. Within 0.005 of its peak, iECR is flat over $\lambda \in [1.29, 1.50]$, $\lambda_{\%} \in [0.030, 0.044]$, $ES \in [0.93, 1.07]$, and $\rho \in [0.80, 0.86]$. The tuning parameter therefore needs only to land in the right neighborhood. Method C reaches a higher peak than Method A (0.930 vs. 0.908), even though its optimum $ES^* \approx 1.00$ sits below $\lambda^* \approx 1.43$.

6 | DISCUSSION

Summary of findings.

Across the case-study grid all four methods reach broadly similar iECR, with Method A highest. The notes below are specific to these prespecified settings and this grid, not a general ranking.

- **Method A (σ -Scaled):** most robust general-purpose choice here (highest mean iECR, 0.872), holding across sample

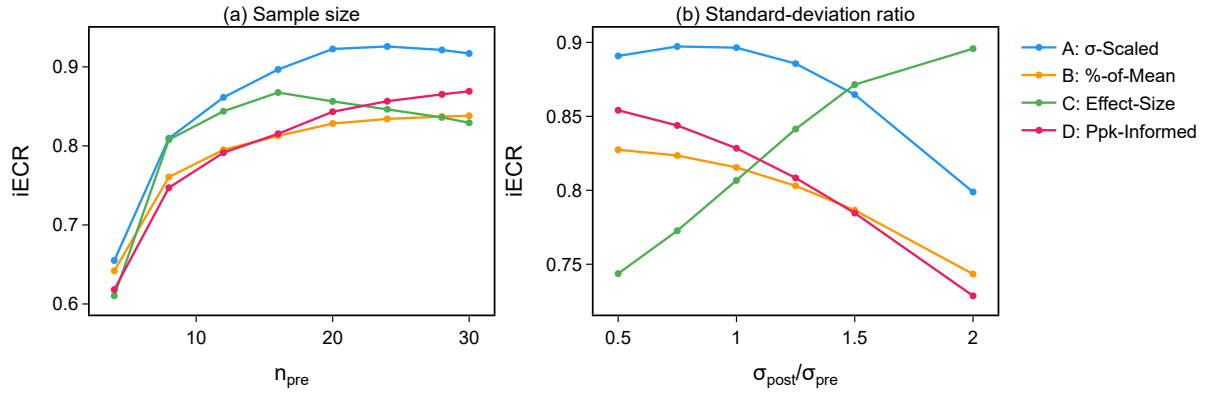


FIGURE 5 Mean iECR for the four methods. (a) against pre-change sample size n_{pre} . (b) against the standard-deviation ratio $\sigma_{post}/\sigma_{pre}$. Both panels come from the same factorial grid as Table 5 ($N_{sim} = 5,000$, $N_{\delta} = 20$), each marginalized over all remaining design factors.

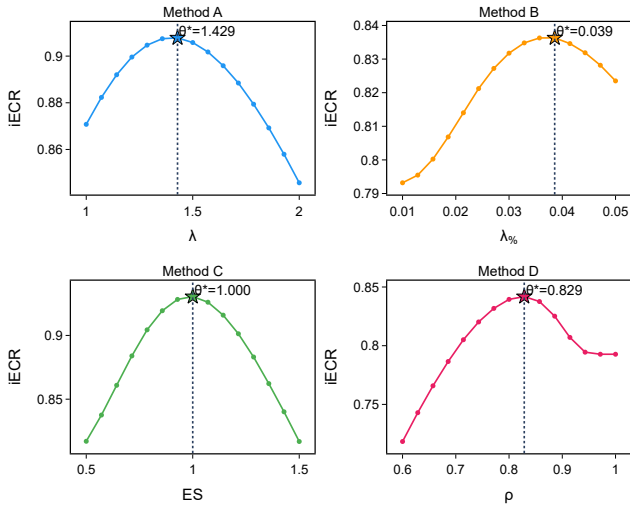


FIGURE 6 Fine-tuned iECR vs. primary tuning parameter, evaluated via Monte Carlo simulation over a factorial subset ($n_{pre} \in \{8, 20, 30\}$, $\sigma_{post}/\sigma_{pre} = 1$, $\kappa = 0.05$; 18 $m/\sigma \times S/\sigma$ combinations per n_{pre} ; fine grid of 15 values per method). Stars and dotted lines mark the iECR-maximizing value θ^* .

sizes and specification widths, though it degrades under variance inflation as B and D do.

- **Method C (Effect-Size):** best under strong variance inflation—the only method that improves as $\sigma_{post}/\sigma_{pre}$ grows; its high default paradox rate subsides once ES is tuned, attaining the highest peak iECR (0.930).
- **Method D (Ppk-Informed):** tightest, capability-tied margin—lowest paradox rate (mean iPR = 0.020) at the cost of the widest gray zone, and needs larger samples.

- **Method B (Percent-of-Mean):** simplest to compute, but its mean-scaled margin swings between paradox-prone (large m/σ) and over-tight (small m/σ), which places it at the lower end of this grid; check the margin it implies at the process's own μ_{pre}/σ_{pre} before adopting it.

Within the tuning subset, tuning each method to its own optimum does *not* make performance converge—a ≈ 9 -point iECR gap persists (Section 5.4)—so the differences appear structural rather than artifacts of the defaults.

Decision framework.

Figure 7 guides method selection. Its first branch is the margin's *scientific purpose*: capability preservation (D), variability scaling (A), or effect-size reasoning (C). That purpose, not a numerical threshold, is what must be defensible; variance assumptions, sample size, and feasibility then refine it. Feasibility deserves care. Under a relative target $Ppk_{target} = \rho Ppk_{pre}$ the condition $EAC_D > 0$ reduces to $\rho < 1$ and always holds. A *fixed* target such as 1.33 instead collapses the margin at the boundary (Appendix G). If a fixed target is infeasible, relax it rather than switching method, which would merely set the requirement aside.

Generality of the ECR framework.

The framework is agnostic to both the EAC method and the domain. Any candidate margin—Bayesian, data-dependent²⁵, or entirely novel—can be evaluated by its ECR, PR, and GZR curves, so the four methods here are a *case study*, not an exhaustive list. It applies to process comparability, method and site transfer, and scale-up, and extends further—from biosimilarity tests³⁵ to economics³⁶. Because PR and GZR are opposing consumer's- and producer's-risk axes, methods can be ranked

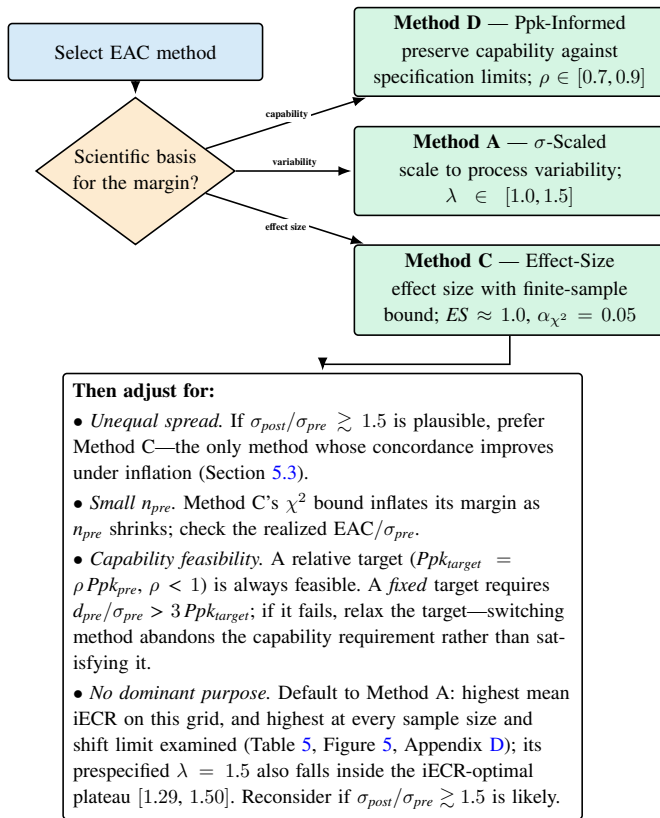


FIGURE 7 Decision framework for EAC method selection. The margin's scientific purpose selects the method; variance assumptions, sample size, and capability feasibility then refine the choice. Method B is not shown (Section 5.2).

by whichever risk dominates¹⁸. Two uses should not be conflated. *Planning* computes the curves *before* a confirmatory study, to justify a margin. The tuning sweep (Section 5.4) can also be run at a process's own pre-change parameters to set its tuning constant, provided the margin is prespecified and scientifically defensible. *Diagnostic benchmarking* (Appendix H) scores an incumbent margin for an observed process; it is exploratory only, and never grounds for retrospectively changing a pre-specified margin.

Limitations.

All methods assume normality. Method D also requires specification limits. The margin is fixed and known, so the rates condition on the pre-change parameters. Appendix F relaxes this and the comparison holds. The iECR is a scalar, so it can mask crossing curves. Report it with the curve (Section 2.3), not instead of it. Results depend on the grid. Its sample-size range ($n_{pre} \in [4, 30]$) reflects pharmaceutical practice and may not generalize. At large n the point-null D-test is overpowered, which inflates the paradox rate.

Future work.

Natural extensions include priors on the shift magnitude (a Bayesian EAC), adaptive re-estimation of σ from incoming data, non-normal ECR curves, and simultaneous testing across correlated attributes. A further extension replaces the point-null D-test with a minimum-effect gate^{37,38}. A difference would be declared only when the test is significant *and* the observed shift exceeds a fraction of the margin (e.g. 0.75Δ). This suppresses the spurious large- n paradoxes while preserving the zone classification.

7 | CONCLUSION

The Equivalence Concordance Rate framework is a quantitative tool for comparing and calibrating equivalence acceptance criteria. It cross-classifies the joint outcome of a difference test and a TOST across simulated scenarios. The ECR, PR, and GZR then reveal whether a margin is wide, narrow, or well-calibrated, and the integrated ECR ranks competing methods directly.

As a case study—not to crown a single best margin—we applied it to four common EAC methods across 10,368 dimensionless scenarios. The methods separated structurally rather than by tuning, with sample size the dominant driver (Appendix B). We recommend the ECR framework as a complementary diagnostic—alongside traditional operating-characteristic analyses—for selecting and validating EAC methods in pharmaceutical manufacturing, and for tuning a prespecified margin to a specific process from its pre-change data.

AUTHOR CONTRIBUTIONS

All authors contributed to the conceptualization and methodology of the study. Emmanuel Atindama wrote the original draft of the manuscript. Huruy Asfha, Qing Meng, and Padraic Nagle contributed to the experimental analysis and to the review and editing of the manuscript. Padraic Nagle is the principal investigator. All authors read and approved the final manuscript.

DATA AVAILABILITY STATEMENT

The comparability measurements analyzed in Appendix H are de-identified manufacturing release data and are proprietary; they are not publicly available. Code implementing the ECR framework is available from the corresponding author upon reasonable request.

USE OF ARTIFICIAL INTELLIGENCE

The authors used AI tools for language editing and to assist with drafting and debugging the simulation code. The scientific ideas, experimental design, and mathematical proofs were conceived and developed by the authors. The authors independently reviewed and validated all manuscript text, references, code, numerical results.

REFERENCES

- Schuurmann DJ. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics* 1987; 15(6): 657–680. doi: 10.1007/BF01068419
- Westlake WJ. Use of confidence intervals in analysis of comparative bioavailability trials. *Journal of Pharmaceutical Sciences* 1972; 61(8): 1340–1341. doi: 10.1002/jps.2600610845
- Wellek S. *Testing Statistical Hypotheses of Equivalence and Noninferiority*. CRC Press. 2nd ed. 2010.
- FDA. Comparability Protocols for Postapproval Changes to the Chemistry, Manufacturing, and Controls Information in an NDA, ANDA, or BLA: Guidance for Industry. U.S. Food and Drug Administration, Center for Drug Evaluation and Research (CDER) and Center for Biologics Evaluation and Research (CBER); 2022. Final guidance; supersedes the 2003 draft.
- ICH. Q5E: Comparability of biotechnological/biological products subject to changes in their manufacturing process. International Council for Harmonisation; 2004.
- Limentani GB, Ringo MC, Ye F, Bergquist ML, McSorley EO. Beyond the *t*-test: Statistical equivalence testing. *Analytical Chemistry* 2005; 77(11): 221A–226A. doi: 10.1021/ac053390m
- Stangler T, Schiestl M. Similarity assessment of quality attributes of biological medicines: the calculation of operating characteristics to compare different statistical approaches. *AAPS Open* 2019; 5: 4. doi: 10.1186/s41120-019-0033-9
- Yassouridis C, Dürauer A, Scharl T, Leisch F, Brocard C, Tscheliessnig AL. Strategy to equivalence testing for development and scale up of biopharmaceutical downstream processes. *Chemical Engineering Science* 2021; 235: 116497. doi: 10.1016/j.ces.2021.116497
- EMA. Reflection paper on statistical methodology for the comparative assessment of quality attributes in drug development. European Medicines Agency; 2021. EMA/CHMP/138502/2017.
- Yu B, Zeng L, Yang H. A risk-based approach to setting acceptance criteria for pharmaceutical process comparability. *Journal of Pharmaceutical Innovation* 2017; 12(4): 357–366. doi: 10.1007/s12247-017-9295-8
- Yu B, Zeng L, Yang H. Determination of acceptance criteria and sample sizes for accelerated stability comparability studies for biologics. *Biologics* 2017; 49: 46–50. doi: 10.1016/j.biologics.2017.07.001
- Chatfield MJ, Borman PJ. Acceptance criteria for method equivalency assessments. *Analytical Chemistry* 2009; 81(24): 9841–9848. doi: 10.1021/ac901944t
- Chatfield MJ, Borman PJ, Damjanov I. Evaluating change during pharmaceutical product development and manufacture — comparability and equivalence. *Quality and Reliability Engineering International* 2011; 27(5): 629–640. doi: 10.1002/qre.1230
- Little TA. Equivalence testing for comparability. *BioPharm International* 2015; 28(2): 45–48.
- Burdick RK, Sidor L. Establishment of an equivalence acceptance criterion for accelerated stability studies. *Journal of Biopharmaceutical Statistics* 2013; 23(4): 730–743. doi: 10.1080/10543406.2013.789891
- Burdick RK, LeBlond DJ, Pfahler LB, et al. *Statistical Applications for Chemistry, Manufacturing and Controls (CMC) in the Pharmaceutical Industry*. Cham: Springer International Publishing. 2017. <https://doi.org/10.1007/978-3-319-50186-4>.
- Goeman JJ, Solari A, Stijnen T. Three-sided hypothesis testing: Simultaneous testing of superiority, equivalence and inferiority. *Statistics in Medicine* 2010; 29(20): 2117–2125. doi: 10.1002/sim.4002
- Lakens D. Equivalence tests: A practical primer for *t* tests, correlations, and meta-analyses. *Social Psychological and Personality Science* 2017; 8(4): 355–362. doi: 10.1177/1948550617697177
- Lakens D, Scheel AM, Isager PM. Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science* 2018; 1(2): 259–269. doi: 10.1177/2515245918770963
- Isager PM, Fitzgerald J. Three-sided testing to establish practical significance: A tutorial. Tinbergen Institute Discussion Paper No. TI 2024-077/III; 2024.
- Campbell H, Gustafson P. Conditional equivalence testing: An alternative remedy for publication bias. *PLOS ONE* 2018; 13(4): e0195145. doi: 10.1371/journal.pone.0195145
- Blume JD, D'Agostino McGowan L, Dupont WD, Greevy RA. Second-generation *p*-values: Improved rigor, reproducibility, & transparency in statistical analyses. *PLOS ONE* 2018; 13(3): e0188299. doi: 10.1371/journal.pone.0188299
- Blume JD, Greevy RA, Welty VF, Smith JR, Dupont WD. An introduction to second-generation *p*-values. *The American Statistician* 2019; 73(sup1): 157–167. doi: 10.1080/00031305.2018.1537893
- Lakens D, Delacre M. Equivalence testing and the second generation *p*-value. *Meta-Psychology* 2020; 4: MP.2018.933. doi: 10.15626/MP.2018.933
- Koobs S, Koning NW. Equivalence testing with data-dependent and post-hoc equivalence margins. arXiv preprint arXiv:2603.16213; 2026.
- Kruschke JK. Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science* 2018; 1(2): 270–280. doi: 10.1177/2515245918771304
- Kruschke JK. Equivalence testing (two one-sided test) and NHST compared with HDI and ROPE [Blog post]. *Doing Bayesian Data Analysis*; 2017. <https://doingbayesiandataanalysis.blogspot.com/2017/02/equivalence-testing-two-one-sided-test.html>.
- Linde M, Tendeiro JN, Selker R, Wagenmakers EJ, van Ravenzwaaij D. Decisions about equivalence: A comparison of TOST, HDI-ROPE, and the Bayes factor. *Psychological Methods* 2023; 28(3): 740–755. doi: 10.1037/met0000402
- Dienes Z, McLatchie N. Four reasons to prefer Bayesian analyses over significance testing. *Psychonomic Bulletin & Review* 2018; 25(1): 207–218. doi: 10.3758/s13423-017-1266-z
- Montgomery DC. *Introduction to Statistical Quality Control*. Wiley. 8th ed. 2019.
- Wang T, Chow SC. On the establishment of equivalence acceptance criterion in analytical similarity assessment. *Journal of Biopharmaceutical Statistics* 2017; 27(2): 206–212. doi: 10.1080/10543406.2016.1265539
- ICH. M13A: Bioequivalence for immediate-release solid oral dosage forms. International Council for Harmonisation; 2024. Harmonised guideline, Step 4 document, adopted 23 July 2024.
- Boulaguiem Y, Quartier J, Lapteva M, et al. Finite sample corrections for average equivalence testing. *Statistics in Medicine* 2024; 43(5): 833–854. doi: 10.1002/sim.9993
- Phillips KF. Power of the two one-sided tests procedure in bioequivalence. *Journal of Pharmacokinetics and Biopharmaceutics* 1990; 18(2): 137–144. doi: 10.1007/BF01063556
- Zahel T. A novel bootstrapping test for analytical biosimilarity. *The AAPS Journal* 2022; 24(6): 112. doi: 10.1208/s12248-022-00749-3
- Fitzgerald J. The need for equivalence testing in economics. I4R Discussion Paper Series No. 125; 2024.

37. Serlin RC, Lapsley DK. Rationality in psychological research: The good-enough principle. *American Psychologist* 1985; 40(1): 73–83. doi: 10.1037/0003-066X.40.1.73
38. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates. 2nd ed. 1988.

APPENDIX

A PROOFS

This appendix proves Proposition 1 (centered case) and Corollary 1 (off-center correction).

Proof of Proposition 1. Let $\mu_{pre} = \frac{USL+LSL}{2}$ be the pre-change mean. Let $S = \frac{USL-LSL}{2}$ be the specification half-width. From (10),

$$Ppk_{pre} = \min\left(\frac{USL - \mu_{pre}}{3\sigma_{pre}}, \frac{\mu_{pre} - LSL}{3\sigma_{pre}}\right) \quad (A1)$$

$$= \min\left(\frac{\frac{USL-LSL}{2}}{3\sigma_{pre}}, \frac{\frac{USL-LSL}{2}}{3\sigma_{pre}}\right) = \frac{S}{3\sigma_{pre}}. \quad (A2)$$

Now shift the mean by δ , so $\mu_{post} = \mu_{pre} + \delta$. The distances to the limits become $S - \delta$ and $S + \delta$. Apply (10) at μ_{post} , with $\sigma_{post} = \sigma_{pre}$ by hypothesis:

$$Ppk_{post} = \min\left(\frac{S - \delta}{3\sigma_{pre}}, \frac{S + \delta}{3\sigma_{pre}}\right) = \frac{S - |\delta|}{3\sigma_{pre}}. \quad (A3)$$

Impose the capability constraint $Ppk_{post} \geq Ppk_{target}$:

$$\frac{S - |\delta|}{3\sigma_{pre}} \geq Ppk_{target} \implies |\delta| \leq S - 3 \cdot Ppk_{target} \cdot \sigma_{pre}. \quad (A4)$$

If $3 \cdot Ppk_{target} \cdot \sigma_{pre} > S$, the right-hand side is negative. The target is then infeasible even at zero shift. Clamping there yields (11). $\square$

Proof of Corollary 1. Off-center, the two distances to the limits are unequal. The symmetric argument of Proposition 1 must therefore be replaced by a worst-direction one. Write $a = USL - \mu_{pre}$ and $b = \mu_{pre} - LSL$, and set $d_{pre} = \min(a, b)$. Since $a + b = 2S$,

$$d_{pre} = \min(a, b) \leq \frac{a+b}{2} = S, \quad (A5)$$

with equality if and only if $a = b$, i.e. the process is centered. Shift the mean by δ . The distances become $a - \delta$ and $b + \delta$. Again with $\sigma_{post} = \sigma_{pre}$,

$$Ppk_{post} = \frac{\min(a - \delta, b + \delta)}{3\sigma_{pre}}. \quad (A6)$$

For any δ , $a - \delta \geq a - |\delta| \geq d_{pre} - |\delta|$ and $b + \delta \geq b - |\delta| \geq d_{pre} - |\delta|$, hence

$$\min(a - \delta, b + \delta) \geq d_{pre} - |\delta|, \quad (A7)$$

Equality holds when the shift points toward the nearer limit. That is the binding case. Requiring $Ppk_{post} \geq Ppk_{target}$ for every $|\delta| \leq \Delta$ therefore reduces to

$$\frac{d_{pre} - \Delta}{3\sigma_{pre}} \geq Ppk_{target} \implies \Delta \leq d_{pre} - 3Ppk_{target}\sigma_{pre}. \quad (A8)$$

Clamping at zero yields the robust margin (12). $\square$

B ANOVA OF FACTORIAL DESIGN FACTORS

Table B1 reports each design factor's contribution to iECR, from a Type II ANOVA of the factorial results of Section 5.2:

$$\begin{aligned} \text{iECR} = & \beta_0 + \text{Method} + \sum_j \beta_j x_j + \\ & \sum_j \gamma_j (\text{Method} \times x_j) + \varepsilon, \end{aligned} \quad (B1)$$

where the design factors x_j are n_{pre} , the sample-size ratio $r = n_{post}/n_{pre}$, $\sigma_{post}/\sigma_{pre}$, S/σ_{pre} , m/σ_{pre} , and κ . Type II sums of squares are used, since the constraint $4 \leq n_{post} \leq 30$ leaves the grid unbalanced. Effect sizes are $\eta^2 = SS_{\text{factor}}/SS_{\text{total}}$, read against Cohen's benchmarks³⁸: 0.01 small, 0.06 medium, 0.14 large. Section 5.3 interprets the leading effects.

Four terms fall below the small-effect threshold and are omitted: κ , $M \times \kappa$, S/σ_{pre} , and r , all with $\eta^2 \leq 0.009$. The ratio still acts through its interaction with method, the fourth-largest effect. The grid is exhaustive and method-scenario rows are paired, so p -values carry no sampling interpretation; effect magnitudes are reported instead.

TABLE B1 Contribution of each design factor to iECR. Factors with negligible effect are excluded. M = Method, r = sample-size ratio n_{post}/n_{pre} .

Factor	SS	F	η^2	Effect
n_{pre}	144.10	5,332	0.290	Large
$M \times \sigma_{post}/\sigma_{pre}$	58.88	1,017	0.118	Medium
M	32.81	2,833	0.066	Medium
$M \times r$	27.98	805	0.056	Small
$M \times n_{pre}$	21.63	267	0.043	Small
$M \times m/\sigma_{pre}$	16.20	280	0.033	Small
$M \times S/\sigma_{pre}$	13.04	563	0.026	Small
$\sigma_{post}/\sigma_{pre}$	10.00	518	0.020	Small
m/σ_{pre}	5.39	279	0.011	Small
Residual	159.70	—	0.321	—

C OVERPOWERED DIFFERENCE TEST AT LARGE N

As the sample size grows, the point-null D-test detects ever-smaller shifts. Those shifts are practically negligible. The paradox rate (Z_3) therefore rises (Figure C1). This is the mechanism behind the large- n turnover in Methods A and C (Section 5.3). It is also the practical-significance caveat noted in the Limitations. The rise holds for Methods A, B, and D. For Method C the integrated rate is U-shaped, with a minimum near $n_{pre} = 16$. At a fixed shift of $0.5\sigma_{pre}$ every method approaches a paradox rate of 1 by $n = 200$, Method D last.

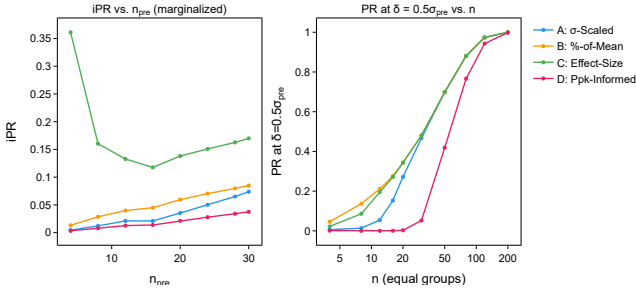


FIGURE C1 Paradox rate against sample size for the four methods. Left: integrated paradox rate vs. n_{pre} , marginalized over all other design factors. Right: paradox rate at a fixed shift $\delta = 0.5\sigma_{pre}$, with equal groups from $n = 4$ to 200 at the baseline parameterization.

D IECDR SHIFT-RANGE SENSITIVITY

The iECR averages $ECR(\delta)$ over a finite shift range $[0, \delta_{max}]$. The paper reports two rankings: one at the baseline (Table 4) and one across the grid (Table 5). We check whether either is an artifact of that choice. Neither is.

Figure D1 varies δ_{max} at the baseline scenario: $m/\sigma_{pre} = 50$, $S/\sigma_{pre} = 5$, $n_{pre} = n_{post} = 20$, $\sigma_{post}/\sigma_{pre} = 1$, and $\kappa = 0.05$. The baseline order $D > A > C > B$ holds for every $\delta_{max}/\sigma_{pre} \in \{2, 3, 4, 5, 6\}$. Table D1 repeats the check over 648 grid scenarios. The subset is stratified rather than a uniform subsample: it takes $n_{pre} \in \{4, 8, 12, 20, 30\}$ with $r \in \{1/2, 1, 2\}$, three m/σ_{pre} levels, and $\kappa = 0.05$, so it omits the smallest sample-size ratio $r = 1/3$. There A is highest at every limit, as in the full-grid comparison. Widening the range raises every absolute level and draws the methods together, narrowing their spread from 0.113 to 0.042. C and D come within under 0.002 of each other and cross on the way, a gap too small for ranking.

The middle pair is sensitive to the scenario mix as well as to the limit. At $\delta_{max} = 4\sigma_{pre}$ this subset puts D above C (0.786 vs 0.784), while the full grid at that same limit puts C above D

(0.822 vs 0.808, Table 5). With the limit held fixed, the C–D order tracks which scenarios are averaged, not the shift range. Scope explains the baseline–grid difference too: one is a single scenario, the other a 648-scenario average, as Tables 4 and 5 already show. We adopt $\delta_{max} = 4\sigma_{pre}$.

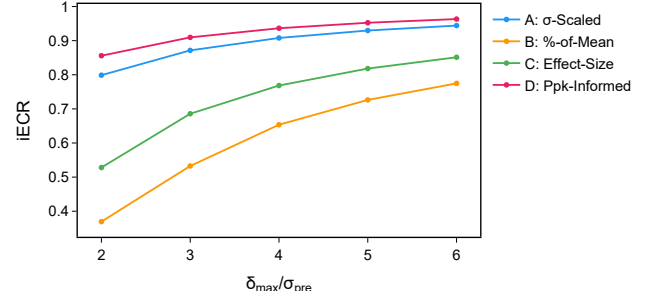


FIGURE D1 iECR versus the shift-range upper bound $\delta_{max}/\sigma_{pre}$, at the single baseline scenario ($N_{sim} = 5,000$, 60 shifts).

TABLE D1 Mean iECR at each shift limit $\delta_{max}/\sigma_{pre}$, across 648 grid scenarios ($N_{sim} = 2,000$, 25 shifts).

$\delta_{max}/\sigma_{pre}$	A	B	C	D	Ranking
2	0.706	0.593	0.675	0.604	$A > C > D > B$
3	0.787	0.704	0.732	0.718	$A > C > D > B$
4	0.837	0.775	0.784	0.786	$A > D > C > B$
5	0.869	0.819	0.826	0.828	$A > D > C > B$
6	0.891	0.849	0.854	0.857	$A > D > C > B$

E UNEQUAL-VARIANCE ROBUSTNESS: WELCH-SATTERTHWAITE CHECK

Both tests pool variances throughout ($df = n_{pre} + n_{post} - 2$), which is misspecified when $\sigma_{post} \neq \sigma_{pre}$. We therefore repeat the marginal $\sigma_{post}/\sigma_{pre}$ analysis with Welch–Satterthwaite versions of both tests, over a representative grid subset. Table E1 shows the result for Methods A and C, which determine the crossover. Welch lowers absolute iECR, but C still overtakes A at $\sigma_{post}/\sigma_{pre} \approx 1.5$. Method C’s advantage under variance inflation is therefore genuine, not an artifact of pooling.

F UNCONDITIONAL MARGINS

The main analysis fixes each margin at its population value (Section 4). Here the full grid experiment is repeated with every margin re-estimated from each replicate’s own simulated pre-change sample. Seeds and simulated datasets are those of Table 5; only the margin rule differs.

TABLE E1 Mean iECR for Methods A and C against the standard-deviation ratio, under the pooled and Welch procedures. Representative grid subset, $N_{sim} = 2,000$.

$\sigma_{post}/\sigma_{pre}$	Pooled		Welch	
	A	C	A	C
0.50	0.879	0.734	0.877	0.724
0.75	0.891	0.765	0.885	0.761
1.00	0.893	0.800	0.883	0.797
1.25	0.884	0.836	0.867	0.825
1.50	0.865	0.867	0.837	0.841
2.00	0.800	0.891	0.749	0.830

TABLE F1 Method comparison with unconditional margins. Columns, grid, and simulation settings are as in Table 5.

Method	Mean iECR	SD	min iECR	iGZR	iPR	max GZR	max PR	EAC/ σ
A: σ -Scaled	0.874	0.109	0.395	0.094	0.032	0.225	0.146	1.47
C: Effect-Size	0.831	0.087	0.488	0.019	0.150	0.036	0.447	2.28
D: Ppk-Informed	0.807	0.111	0.381	0.173	0.020	0.601	0.086	0.96
B: %-of-Mean	0.800	0.112	0.381	0.146	0.054	0.500	0.170	1.22

The ranking is unchanged from Table 5. Only Method C moves materially (mean iECR $0.822 \rightarrow 0.831$, minimum $0.337 \rightarrow 0.488$), as expected: the conditional design leaves its χ^2 hedge against an underestimated σ_{pre} unexercised. The gain is largest at small n_{pre} ($+0.048$ at $n_{pre} = 4$, $+0.003$ at 30). Methods A, B, and D shift by less than 0.002.

G METHOD D FEASIBILITY CLIFF UNDER A FIXED CAPABILITY TARGET

The main design uses a relative target, $Ppk_{target} = \rho Ppk_{pre}$. Consider instead a fixed target such as $Ppk_{target} = 1.33$. For a centered process the margin is then zero until $S/\sigma_{pre} = 3 Ppk_{target} \approx 4$, and positive only beyond it (Figure G1). Below the cliff no usable margin exists, since a zero margin can never declare equivalence. We therefore report iECR only in the feasible region. There it peaks near $S/\sigma_{pre} = 5.25$, then declines as the widening margin admits paradoxes. The relative target ($\rho = 0.8$) scales with actual capability and avoids the cliff (Section 6).

H APPLICATION TO REAL DATA: A WORKED EXAMPLE

This appendix illustrates how the framework compares a user-defined margin against a recommended one. We use three attributes of a real bulk-versus-packaged comparability dataset (Table H1). For each attribute the ECR diagnostics recommend

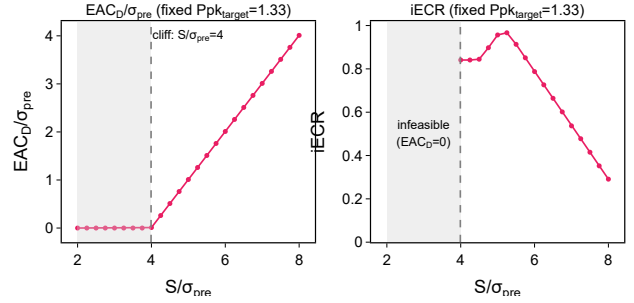


FIGURE G1 Method D under a fixed target $Ppk_{target} = 1.33$, against the specification ratio S/σ_{pre} . Left: EAC_D/σ_{pre} . Right: iECR, plotted only where $EAC_D > 0$. The dashed line marks $S/\sigma_{pre} = 3 Ppk_{target} \approx 4$ and the gray region is infeasible ($EAC_D = 0$). Centered process, $n_{pre} = n_{post} = 20$, $N_{sim} = 3,000$, 15 shifts.

TABLE H1 Summary statistics for the three comparability attributes. Bulk is the pre-change group, packaged the post-change group.

Attribute	Group	N	Mean	SD
Assay	Bulk	21	99.2	1.38
	Packaged	9	99.4	1.11
Impurities	Bulk	21	0.33	0.08
	Packaged	10	0.23	0.13
Water	Bulk	20	1.32	0.12
	Packaged	10	1.33	0.17

TABLE H2 * marks the higher-iECR margin. Δ is the absolute margin in data units; the calibration profile (iECR, min ECR, max GZR, max PR) is simulated over $[0, 4\sigma_{pre}]$ at each attribute's estimated parameters.

Attribute	Margin	Δ	iECR	min ECR	max GZR	max PR
Assay	Effect-Size*	1.881	0.961	0.824	0.114	0.062
	User-Defined	4.140	0.557	0.011	0.000	0.990
Impurities	Ppk-Informed	0.283	0.971	0.909	0.007	0.084
	User-Defined*	0.240	0.977	0.936	0.041	0.023
Water	σ -Scaled*	0.171	0.934	0.791	0.194	0.017
	User-Defined	0.360	0.663	0.189	0.000	0.811

one of the four EAC methods (Section 5.2). We then profile that margin against a fixed $3 s_{pre}$ user-defined margin (Table H2).

For assay and water the recommended margin is clearly better on iECR (0.961 vs. 0.557, and 0.934 vs. 0.663). For impurities the two are close (0.971 vs. 0.977). The choice then rests on which error matters more. The Ppk-Informed margin gives the lower gray-zone rate (max GZR = 0.007); the user-defined margin gives the lower paradox rate (max PR = 0.023). Surfacing that trade-off is the point of the diagnostic.