arXiv:2512.16012v1 [stat.ME] 17 Dec 2025

Reliability-Targeted Simulation of Item Response Data: Solving the Inverse Design Problem

JoonHo Lee

Abstract. Monte Carlo simulations are the primary methodology for evaluating Item Response Theory (IRT) methods, yet marginal reliability—the fundamental metric of data informativeness—is rarely treated as an explicit design factor. Unlike in multilevel modeling where the intraclass correlation (ICC) is routinely manipulated, IRT studies typically treat reliability as an incidental outcome, creating a "reliability omission" that obscures the signal-to-noise ratio of generated data. To address this gap, we introduce a principled framework for *reliability-targeted simulation*, transforming reliability from an implicit by-product into a precise input parameter. We formalize the inverse design problem, solving for a global discrimination scaling factor that uniquely achieves a pre-specified target reliability. Two complementary algorithms are proposed: Empirical Quadrature Calibration (EQC) for rapid, deterministic precision, and Stochastic Approximation Calibration (SAC) for rigorous stochastic estimation. A comprehensive validation study across 960 conditions demonstrates that EQC achieves essentially exact calibration, while SAC remains unbiased across non-normal latent distributions and empirical item pools. Furthermore, we clarify the theoretical distinction between average-information and error-variance-based reliability metrics, showing they require different calibration scales due to Jensen's inequality. An accompanying open-source R package, `IRTsimrel`, enables researchers to standardize reliability as a controlled experimental input.

Keywords: Item Response Theory; reliability-targeted simulation; inverse design problem; marginal reliability; Empirical Quadrature Calibration; Stochastic Approximation Calibration

Date: December 19, 2025.

Lee: Department of Educational Studies in Psychology, Research Methodology, and Counseling, The University of Alabama, jlee296@ua.edu.

Acknowledgments: The author is grateful for the support of the Institute of Education Sciences Grant R305D240078.

1. Introduction

Monte Carlo simulation studies constitute the primary methodology for evaluating Item Response Theory (IRT) estimation methods, model selection criteria, and scoring algorithms. In the absence of "ground truth" in empirical data, simulations provide the sole environment where latent parameters are known, allowing researchers to quantify bias, efficiency, and error rates (Morris et al., 2019). Consequently, the rigor of a simulation design directly determines the validity and generalizability of its findings. Standard practice in IRT simulation involves the systematic manipulation of sample size, test length, item parameter distributions, and latent distribution shapes (e.g., Cheng & Meng, 2025; Guastadisegni et al., 2025; Monroe & Cai, 2014; Paganin et al., 2022; Woods & Thissen, 2006). Researchers meticulously cross these factors to explore how estimators behave under varying conditions of data scarcity and structural complexity.

Yet a systematic methodological gap persists: *marginal reliability*—the fundamental metric of data informativeness and signal-to-noise ratio (Brennan & Kane, 1977; Cronbach & Gleser, 1964; Guyatt et al., 1992; Rouder & Mehrvarz, 2024)—is rarely treated as an explicit design factor. Instead, reliability is almost invariably treated as an implicit outcome of item parameter selection, reported sporadically if at all. This "reliability omission" creates a significant blind spot. Readers are often left unable to determine the informational quality of the generated data. Without explicit control or reporting of marginal reliability, it remains unclear whether a proposed method's performance is robust across the spectrum of precision found in operational testing, or if it is contingent upon a specific, potentially idealized, high-reliability regime.

This omission persists even in rigorous, high-quality methodological studies. For instance, Paganin et al. (2022) conducted a sophisticated evaluation of Bayesian semiparametric IRT models, varying sample sizes, test lengths, and complex multimodal latent distributions. Despite this attention to distributional nuance, the marginal reliability of the generated datasets was neither controlled nor reported. Similarly, recent studies investigating non-normal latent traits (e.g., Bambirra Gonçalves et al., 2018; Cheng & Meng, 2025; Guastadisegni et al., 2025; Wang et al., 2018) vary structural parameters but leave the resulting signal-to-noise ratio uncontrolled. These cases exemplify a broader disciplinary tendency: while the structure of simulated data is rigorously engineered, its informational strength is frequently left uncalibrated.

The omission of reliability in IRT simulations stands in stark contrast to the methodological practice of the cognate field of multilevel modeling (MLM). In MLM

simulation methodology, the intraclass correlation (ICC) is systematically treated as a primary design factor (Maas & Hox, 2005; McNeish & Stapleton, 2016). This practice reflects the structural isomorphism between the two metrics: just as the ICC quantifies the proportion of variance attributable to between-group differences (signal) relative to residual error (noise), marginal reliability quantifies the proportion of variance attributable to the latent trait relative to measurement error. Both metrics share the same mathematical structure—a Rayleigh quotient representing projection efficiency in Hilbert space (cf. Zumbo, 2025, for the Hilbert space formalization of reliability)—which explains their analogous roles as design factors in their respective domains.

In MLM research, it is standard practice to explicitly vary the ICC across conditions (e.g., $\rho \in \{0.10, 0.20, 0.30\}$) to examine method performance across different informativeness regimes (e.g., Can et al., 2015; Cho et al., 2015; Hsu et al., 2017; Lüdtke et al., 2017). This design choice reflects a consensus that the ratio of signal variance to error variance is a fundamental property of the data structure that dictates when model complexity is warranted and how estimators perform (Lee & Wind, 2025; Lee et al., 2025; Paddock et al., 2006). By contrast, IRT simulation practice has not yet internalized reliability as a first-class design factor. While MLM researchers would rarely simulate clustered data without knowing the ICC, IRT researchers routinely simulate response data without targeting—or often even knowing—the marginal reliability.

The failure to control reliability in simulation design has direct and substantive consequences for the ecological validity of psychometric research. First, it threatens the generalizability of findings to real-world contexts. Operational assessments vary wildly in precision (Ramsay, 2016), from high-stakes exams with reliabilities exceeding 0.90 to short formative assessments or screeners where reliability may hover between 0.50 and 0.70 (Conoyer et al., 2022; Frisbie, 1988). Simulations that implicitly generate data with consistently high reliability may overstate the utility of complex models or underestimate the fragility of estimators in "messy," low-information environments—such as clinical settings with small samples ($N = 115$) and high missingness rates (e.g., 42%) (Gilholm et al., 2021).

Second, uncontrolled reliability acts as a confound in model comparison. Conclusions about the superiority of one estimator over another may be contingent on the underlying reliability regime. For example, Soland et al. (2024) demonstrated that scoring decisions interact with test reliability to alter study conclusions, finding that

Bayesian shrinkage methods (EAP) produced valid inference in high-reliability conditions but led to massive inflation of Type I error rates in low-reliability settings. If a simulation study defaults to a high-reliability condition without explicit control, such pathologies may remain undetected, leading to an incomplete understanding of an estimator's properties. Finally, the lack of reliability reporting limits cumulative science; without knowing the achieved reliability of simulated datasets, it is difficult to meaningfully compare performance metrics (e.g., bias, RMSE) across different studies.

To address this gap, this paper introduces a principled framework for *reliability-targeted simulation* of IRT data. Our primary objective is to transform marginal reliability from an implicit outcome into an explicit input parameter, allowing researchers to generate data that achieves a pre-specified reliability target while preserving realistic distributional characteristics.

The specific contributions of this study are fourfold:

(1) **Mathematical Framework:** We formalize the "inverse design problem" in IRT simulation. We define the relationship between data-generating parameters and marginal reliability, establishing the monotonicity conditions required to uniquely map a global discrimination scaling factor to a target reliability.

(2) **Algorithmic Toolkit:** We propose two complementary calibration algorithms. *Algorithm 1 (Empirical Quadrature Calibration; EQC)* offers a fast, deterministic method suitable for routine use, utilizing large-sample Monte Carlo quadrature. *Algorithm 2 (Stochastic Approximation Calibration; SAC)* employs a Robbins–Monro stochastic approximation approach (Robbins & Monro, 1951) for complex data-generating processes where deterministic quadrature is infeasible.

(3) **Validation Study:** We validate these algorithms across a diverse set of conditions, including non-normal latent distributions (skewed, bimodal, heavy-tailed) and realistic item parameters drawn from the Item Response Warehouse (Domingue et al., 2025; Zhang et al., 2025). We demonstrate that our framework achieves target reliabilities within strict tolerance levels and elucidate the theoretical distinction between average-information and error-variance-based reliability metrics, showing they require different calibration scales due to Jensen's inequality.

(4) **Software Implementation:** We provide an open-source R package, `IRTsimrel`, which implements these algorithms. This tool enables applied researchers to easily integrate reliability targeting into their existing simulation workflows.

The remainder of this paper is organized as follows. Section 2 presents the mathematical framework, defining the measurement models and the specific reliability estimands used. Section 3 details the EQC and SAC calibration algorithms. Section 4 reports the results of a comprehensive validation study comparing the proposed methods against standard benchmarks. Section 5 discusses the implications of reliability-targeted simulation for the field, offers practical recommendations for simulation design, and outlines limitations and future directions.

2. Mathematical Framework

To turn marginal reliability from an incidental outcome of simulation design into an explicit control variable, we must formalize how reliability arises from the underlying item response model and the data-generating process (DGP). This section (a) specifies the generative measurement model, (b) defines the target reliability functionals based on error variance and information, and (c) formulates the inverse design problem in terms of a global discrimination scaling factor.

2.1. Measurement Model and Notation.

Consider a test with I dichotomous items administered to N persons. Let $Y_{pi} \in \{0, 1\}$ denote the response of person p to item i. We employ the two-parameter logistic (2PL) model as the general framework (Birnbaum, 1968), where the probability of a correct response is given by:

$$\Pr(Y_{pi} = 1 \mid \theta_p, \beta_i, \lambda_i) = \text{logit}^{-1}\{\lambda_i(\theta_p - \beta_i)\}. \tag{2.1}$$

Here, θ_p is the latent ability, β_i is the item difficulty, and λ_i is the item discrimination. The Rasch model corresponds to the structural constraint where all discriminations are uniform and fixed at unity ($\lambda_i \equiv 1$) (Rasch, 1980).

In a simulation context, these parameters are treated as random draws from a structural configuration $\Psi = (G, H_\beta, H_\lambda, I)$. Person abilities follow a latent distribution G (typically standardized with variance σ_θ^2). Item difficulties are drawn from H_β, and baseline discriminations from H_λ. The framework accommodates dependence between item parameters, allowing simulations to mimic empirical item pools where difficulty and discrimination are correlated.

2.2. **Test Information and Reliability Definitions.** In classical test theory, reliability is a single number attached to a test. In IRT, measurement precision is *conditional* on the latent trait. The *test information function (TIF)*, assuming local independence, is the sum of item Fisher information functions (Baker & Kim, 2004; de Ayala, 2022; Fischer & Molenaar, 1995):

$$\mathcal{J}(\theta) = \sum_{i=1}^{I} \lambda_i^2 \, \pi_i(\theta)\{1 - \pi_i(\theta)\}, \tag{2.2}$$

where $\pi_i(\theta)$ is the response probability. The asymptotic conditional standard error of measurement is $\mathrm{SEM}(\theta) = 1/\sqrt{\mathcal{J}(\theta)}$. While IRT emphasizes conditional precision, simulation design requires a scalar index of measurement quality. We therefore define two forms of marginal reliability.

MSEM-based Marginal Reliability. Following standard variance decomposition definitions (Thissen & Wainer, 2001; Yang et al., 2012), marginal reliability is the ratio of true variance to total variance. We define the *average error variance*—formally the Mean Squared Error of Measurement (MSEM)—as the expectation of the conditional error variances over the population:

$$\mathrm{MSEM} = \mathbb{E}_{\theta\sim G}\left[\frac{1}{\mathcal{J}(\theta)}\right], \qquad \bar{w} = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \mathrm{MSEM}}. \tag{2.3}$$

This quantity $\bar{w}$ is our primary estimand, representing the reliability of maximum likelihood estimates. Note that MSEM here refers to the mean of the *squared* errors (variances), ensuring dimensional consistency with the variance-ratio formula.

Average-Information Reliability. In design contexts where computational speed is paramount or a summary of "design informativeness" is required, we use a simplified index:

$$\bar{\mathcal{J}} = \mathbb{E}_{\theta\sim G}\big[\mathcal{J}(\theta)\big], \qquad \tilde{\rho} = \frac{\sigma_\theta^2 \, \bar{\mathcal{J}}}{\sigma_\theta^2 \, \bar{\mathcal{J}} + 1}. \tag{2.4}$$

We term $\tilde{\rho}$ the *average-information reliability*. This metric summarizes the heterogeneous information curve $\mathcal{J}(\theta)$ into a single mean value $\bar{\mathcal{J}}$ before applying the variance-ratio transformation.

This formulation parallels the concept of "average reliability" used in multilevel modeling (Lee et al., 2025; Rabe-Hesketh & Skrondal, 2022), where heterogeneous standard errors are condensed into a single summary statistic to guide design decisions. Similarly, it follows the logic of design effects in educational measurement (Adams, 2005), which characterize the reduction in posterior uncertainty relative to the prior. Although test information should not be equated with reliability itself

(Doran, 2005), $\tilde{\rho}$ serves as a valid "reliability-like" summary index that approximates $\bar{w}$ when the TIF is relatively flat over the bulk of the latent distribution.

Note that due to the convexity of $x \mapsto 1/x$, Jensen's inequality implies $\mathbb{E}[1/\mathcal{J}] \geq 1/\mathbb{E}[\mathcal{J}]$, and thus $\tilde{\rho} \geq \bar{w}$. We adopt $\bar{w}$ as the rigorous target for calibration, while retaining $\tilde{\rho}$ as a useful diagnostic of the test's global signal-to-noise potential.

2.3. **The Inverse Design Problem.** Standard psychometric analysis solves a *forward* problem: given a test configuration and a response matrix, estimate item and person parameters and then compute the resulting reliability ρ. Conceptually, if $\rho = f(\Psi)$ denotes the marginal reliability induced by a test configuration $\Psi = (G, H_\beta, H_\lambda, I)$, classical simulation studies simply evaluate $f(\Psi)$ after the fact and, at most, report the realized reliability as a descriptive statistic.

Reliability-targeted simulation reverses this logic. The goal is to solve an *inverse design problem*: Given a target reliability $\rho^* \in (0, 1)$ and a structural configuration Ψ, construct a data-generating mechanism such that the expected marginal reliability of the resulting datasets satisfies $\mathbb{E}[\rho] = \rho^*$.

This inverse problem is inherently non-unique: many combinations of test length, item quality, and latent distribution can yield the same reliability (van der Linden, 2005). A short test with very high discriminations can, in principle, produce the same ρ as a longer test with more modest discriminations. To obtain a tractable and interpretable calibration problem, we identify a *single* control variable that adjusts overall informativeness while preserving the qualitative structure of the test.

We separate the DGP into a structural component and a scalar *scale*:

- **Structure (fixed):** The shape of the latent distribution G, the distribution of item difficulties H_β, the heterogeneity and dependence structure of baseline discriminations $\{\lambda_{i,0}\}$ and their correlation with difficulties, and the test length I.
- **Scale (calibrated):** A global factor $c > 0$ that uniformly rescales all discriminations.

Formally, we define calibrated discriminations by

$$\lambda_i(c) = c\,\lambda_{i,0}, \qquad c > 0,\ i = 1, \ldots, I. \tag{2.5}$$

As c increases, item response curves become steeper, local information $\mathcal{I}_i(\theta)$ grows roughly with c^2 in regions where θ is well aligned with β_i, and the resulting TIF $\mathcal{J}(\theta; c)$ increases on substantial portions of the latent support. Note that if the baseline structure is Rasch ($\lambda_{i,0} \equiv 1$), scaling by c results in a 1PL model with

uniform discrimination c, effectively maintaining the equality of discriminations while adjusting their magnitude.

This parameterization cleanly separates *structure* (how information is distributed across θ) from *scale* (how much total information is available). Adjusting c acts like a "volume knob" on the test's informativeness, changing the signal-to-noise ratio without altering the relative ordering of item qualities or their alignment with G.

Monotonicity of the reliability function. The effectiveness of discrimination scaling as a control variable depends on the behavior of $\rho(c)$. While extremely high discriminations can theoretically degrade MSEM by creating information gaps between items (rendering the function unimodal), within a *practical calibration interval* $[c_L, c_U]$ where the item grid remains sufficiently dense relative to discrimination, the mapping $c \mapsto \rho(c)$ is continuous and strictly increasing. Intuitively, scaling discriminations up within this range uniformly reduces the MSEM, so both $\bar{w}(c)$ and $\tilde{\rho}(c)$ increase monotonically.

A formal statement of the regularity conditions and derivative calculations is provided in Appendix A. For the main text it suffices to record the following consequence:

Corollary 1 (Existence and uniqueness of the calibrated scale)**.** *Let $\rho_{\min} = \rho(c_L)$ and $\rho_{\max} = \rho(c_U)$ denote the reliabilities at the lower and upper calibration bounds. If $\rho(c)$ is continuous and strictly increasing on $[c_L, c_U]$, then for any target $\rho^* \in (\rho_{\min}, \rho_{\max})$ there exists a unique $c^* \in (c_L, c_U)$ such that $\rho(c^*) = \rho^*$.*

Thus, once a calibration interval is chosen, the inverse design problem reduces to solving a well-posed one-dimensional root-finding problem in c. This study develops two algorithms—Empirical Quadrature Calibration (EQC) and Stochastic Approximation Calibration (SAC)—that approximate c^* in practice.

2.4. **Achievable Reliability Bounds.** Even with an optimally chosen scale c^*, not every target reliability is attainable for a fixed test configuration. Structural features of the latent distribution and item pool impose *bounds* on the range of achievable reliabilities. For a given configuration Ψ and calibration interval $[c_L, c_U]$, define

$$\rho_{\min} = \rho(c_L), \qquad \rho_{\max} = \rho(c_U), \qquad \rho_{\min} < \rho^* < \rho_{\max}. \tag{2.6}$$

Any target ρ^* in $(\rho_{\min}, \rho_{\max})$ admits a unique solution c^* as described above; targets outside this range cannot be achieved without altering the underlying structure Ψ or widening the calibration interval.

Several design factors jointly determine $\rho_{\max}$ (and, less dramatically, $\rho_{\min}$):

- **Test length (I).** Holding the item pool fixed, adding items increases information across the trait continuum, raising the ceiling on marginal reliability. For very short tests, even aggressive scaling of discriminations cannot push ρ beyond a moderate level.
- **Item parameter quality.** The distribution of baseline discriminations $\{\lambda_{i,0}\}$ controls how much information can be generated: item pools with many highly discriminating items support larger $\bar{\mathcal{J}}(c)$ and therefore higher potential $\rho_{\max}$, whereas pools dominated by weak items exhibit strong diminishing returns as c increases.
- **Trait–difficulty alignment.** Reliability is highest when item difficulties are well aligned with the bulk of G. If most examinees fall in regions where the TIF is low (e.g., a skewed latent distribution against a nearly symmetric difficulty distribution), the MSEM remains large and $\rho_{\max}$ is depressed, no matter how c is tuned.

In the accompanying `IRTsimrel` R package (see Appendix F for implementation details), attempts to calibrate to a ρ^* outside $(\rho_{\min}, \rho_{\max})$ return a boundary solution (either c_L or c_U) together with a diagnostic warning, signaling that the requested reliability is incompatible with the current test configuration. Appendix B develops the theoretical basis for these feasibility limits and provides practical guidance for checking target achievability; Appendix E illustrates how calibration precision degrades near feasibility boundaries.

3. Calibration Algorithms

Section 2 formalized the reliability-targeted simulation as an inverse design problem: finding a global scaling factor c^* such that the expected marginal reliability of the generated data matches a pre-specified target ρ^*. Formally, we seek the root of the expectation function:

$$\mathbb{E}_\Psi[\rho(c)] - \rho^* = 0. \tag{3.1}$$

Because the reliability function $\rho(c)$ involves complex integrals over latent distributions and item parameters—which may be non-normal or dependent—it generally does not admit a closed-form solution. Consequently, numerical methods must be employed to approximate c^*.

This section details two complementary algorithms for solving this calibration problem: *Empirical Quadrature Calibration (EQC)* and *Stochastic Approximation Calibration (SAC)*. These algorithms are implemented in the accompanying `IRTsimrel`

R package via the `eqc_calibrate()` and `sac_calibrate()` functions, respectively. Both algorithms rely on the realistic data-generating functions described in Appendix C (for latent distributions) and Appendix D (for item parameters via the Item Response Warehouse), treating the generation process as a modular component while focusing on the calibration mechanics.

3.1. **Overview and Design Philosophy.** Our framework relies on a fundamental separation between the *structure* of a test and its *scale*. The *structure* is captured by the configuration $\Psi = (G, H_\beta, H_\lambda, I)$ introduced in Section 2.1—encompassing the latent distribution, item parameter distributions, their dependencies, and test length—and is held fixed throughout calibration. The calibration algorithms operate solely on the *scale* parameter c, adjusting the global discrimination intensity (via $\lambda_i(c) = c \cdot \lambda_{i,0}$) to achieve the target reliability ρ^* without altering the underlying distributional characteristics of the test.

While both algorithms solve the same root-finding problem, they navigate the bias-variance trade-off differently. EQC fixes the Monte Carlo noise to create a smooth deterministic function, making it ideal for routine use. SAC embraces the noise, updating estimates dynamically, which makes it rigorously valid even when fixed quadrature is infeasible. We recommend the following decision rule for applied researchers (Table 1).

TABLE 1. Algorithm Selection Decision Matrix

Scenario	**Recommended**	**Rationale**
Routine simulation work	EQC	High speed, deterministic reproducibility, and negligible error (< 0.01).
Independent validation	SAC (with EQC warm start)	Rigorous verification of EQC solutions against infinite population sampling.
Complex custom DGPs	SAC	Handles dynamic dependencies or stochastic item pools where fixed grids are awkward.
Targeting exact $\bar{w}$	SAC	Targets the MSEM-based parameter directly without quadrature approximation bias.

3.2. **Algorithm 1: Empirical Quadrature Calibration (EQC).** Empirical Quadrature Calibration approximates the reliability function $\rho(c)$ using a large, *fixed* Monte Carlo sample that is reused throughout the calibration process. We refer to this fixed sample as the *empirical quadrature*. Conditional on this sample, the mapping $c \mapsto \hat{\rho}_M(c)$ becomes a smooth, deterministic, and strictly monotonic function. This

transformation allows us to use Brent's method (Brent, 1971), a robust root-finding algorithm, to locate the solution efficiently.

The EQC procedure consists of three stages:

(1) **Construct Empirical Quadrature (Once):** We draw a sample of size M (default $M = 10{,}000$) from the latent distribution G to obtain $\boldsymbol{\theta} = \{\theta_m\}_{m=1}^{M}$. Simultaneously, we draw a single realization of baseline item parameters $\{(\beta_i, \lambda_{i,0})\}_{i=1}^{I}$ from their distribution H. These values remain frozen throughout the calibration.

(2) **Define Empirical Reliability Function:** For any candidate scale c, the calibrated discriminations are $\lambda_i(c) = c \cdot \lambda_{i,0}$. We compute the test information $\mathcal{J}(\theta_m; c)$ for each quadrature point. EQC targets the average-information reliability $\tilde{\rho}$, which ensures computational stability and monotonicity of the objective function.[1] The empirical reliability function is:

$$\hat{\bar{\mathcal{J}}}_M(c) = \frac{1}{M}\sum_{m=1}^{M} \mathcal{J}(\theta_m; c), \qquad \hat{\rho}_M(c) = \frac{\hat{\sigma}_\theta^2\, \hat{\bar{\mathcal{J}}}_M(c)}{\hat{\sigma}_\theta^2\, \hat{\bar{\mathcal{J}}}_M(c) + 1}. \tag{3.2}$$

(3) **Solve via Root-Finding:** Since $\hat{\rho}_M(c)$ is strictly increasing in c, we apply Brent's method to find the unique c^* such that $\hat{\rho}_M(c^*) - \rho^* = 0$ within a tolerance ε.

Algorithm 1: Empirical Quadrature Calibration (EQC)

Input: Target ρ^*, Generators G, H, Quadrature size M, Bounds $[c_L, c_U]$.

Output: Calibrated scale c^*_{EQC}.

(1) **Initialize (Fixed):** Draw $\boldsymbol{\theta}_{1:M} \sim G$ and baseline items $\boldsymbol{\beta}, \boldsymbol{\lambda}_0 \sim H$.

(2) **Define Objective Function** $f(c)$**:**

  (a) Set $\boldsymbol{\lambda} \leftarrow c \cdot \boldsymbol{\lambda}_0$.

  (b) Compute information vector $\mathbf{J} = \mathcal{J}(\boldsymbol{\theta}; \boldsymbol{\beta}, \boldsymbol{\lambda})$.

  (c) Compute $\hat{\rho}_M(c)$ via Equation (3.2).

  (d) Return $\hat{\rho}_M(c) - \rho^*$.

(3) **Solve:** $c^* \leftarrow \text{BrentMethod}(f, \text{bounds} = [c_L, c_U])$.

(4) **Return** c^*.

By the Strong Law of Large Numbers, $\hat{\rho}_M(c)$ converges almost surely to the population reliability $\rho(c)$ as $M \to \infty$, with a Monte Carlo error of order $O(M^{-1/2})$. In

[1] Empirical investigation revealed that the MSEM-based reliability $\bar{w}$ can produce a non-monotone objective under certain item configurations, making root-finding unreliable. EQC is therefore restricted to the average-information metric; users requiring exact $\bar{w}$ targeting should use SAC.

practice, the default $M = 10{,}000$ yields calibration accuracy within ± 0.001, sufficient for most simulation studies. Brent's method ensures superlinear convergence in finding the root, making EQC extremely fast. The default bounds of $c \in [0.3, 3]$ cover most realistic item pools. For typical test lengths (e.g., 20–50 items), this interval generally spans reliability levels from approximately 0.20 to 0.99, encompassing virtually all scenarios of practical psychometric interest. However, these bounds can be adjusted if ρ^* lies near the theoretical boundaries defined in Section 2.4.

3.3. **Algorithm 2: Stochastic Approximation Calibration (SAC).** Stochastic Approximation Calibration addresses scenarios where fixing a quadrature grid is impractical or where exact targeting of the MSEM-based reliability is required. Instead of a fixed sample, SAC employs the Robbins–Monro method (Robbins & Monro, 1951), updating the scale parameter c iteratively using *fresh* Monte Carlo samples at each step (Toulis et al., 2021). This allows the algorithm to integrate over all sources of randomness in the data-generating process.

Let c_n be the scale factor at iteration n. At each step, we draw a fresh batch of data, compute the sample reliability $\hat{\rho}_n$, and update c according to:

$$c_{n+1} = \Pi_{[c_L, c_U]} \left[c_n - a_n(\hat{\rho}_n - \rho^*) \right], \tag{3.3}$$

where Π denotes projection onto the feasible interval, and a_n is a decreasing step size sequence defined by $a_n = a/(n + A)^\gamma$.

To stabilize convergence and ensure asymptotic normality, we employ *Polyak–Ruppert averaging* (Gadat & Panloup, 2023; Polyak & Juditsky, 1992). The final estimator is not the last iterate, but the average of the iterates after a burn-in period B:

$$c^*_{\text{SAC}} = \frac{1}{N - B} \sum_{n=B+1}^{N} c_n. \tag{3.4}$$

Algorithm 2: Stochastic Approximation Calibration (SAC)

Input: Target ρ^*, Iterations N, Burn-in B, Step parameters (a, A, γ).
Output: Calibrated scale c^*_{SAC}.

(1) **Initialize:** Set c_0 (optionally using c^*_{EQC} as warm start).
(2) **Iterate** $(n = 1 \dots N)$**:**
  (a) **Sample:** Draw fresh batch $\boldsymbol{\theta} \sim G$ and items $\Psi \sim H$.
  (b) **Estimate:** Compute $\hat{\rho}_n$ (typically MSEM-based) at scale c_{n-1}.
  (c) **Update:** $c_n \leftarrow c_{n-1} - a_n(\hat{\rho}_n - \rho^*)$ (with projection).

(3) **Average:** Compute mean of $\{c_n\}$ for $n > B$.
(4) **Return** Average.

Users may observe that EQC and SAC yield slightly different calibrated values, with c^*_{SAC} typically exceeding c^*_{EQC} by 5–8%. This is not an algorithmic error but a reflection of the different reliability definitions targeted.

EQC is typically configured to target the *Average Information Reliability* ($\tilde{\rho}$) for computational efficiency, whereas SAC is naturally suited to target the *MSEM-based Reliability* ($\bar{w}$) directly. As established in Section 2.2, Jensen's inequality implies $\tilde{\rho} \geq \bar{w}$ because the harmonic mean of information (used in MSEM) is less than or equal to the arithmetic mean. Consequently, to achieve the *same* numerical target ρ^*, the MSEM-based approach (SAC) requires a higher discrimination scale than the average-information approach (EQC). Ideally, SAC provides the "exact" solution for $\bar{w}$, while EQC provides a close, computationally efficient approximation.

For optimal performance, we recommend using an *EQC warm start*: run EQC first to obtain an approximate solution, and use this value to initialize SAC. This hybrid strategy significantly reduces the burn-in period. Default hyperparameters of $N = 300$, $B = 150$, $a = 1$, $A = 50$, and $\gamma = 0.67$ provide robust convergence across a wide range of latent distributions.

4. Validation Study

This section reports a comprehensive validation study designed to assess whether the proposed calibration algorithms reliably solve the inverse design problem across realistic testing scenarios. The study evaluates (a) calibration accuracy in terms of the deviation between achieved and target reliability, (b) robustness to non-normal latent distributions and empirically realistic item pools, and (c) the theoretical and practical implications of targeting different reliability estimands introduced in Section 2.

4.1. Objectives and Design.

The validation study was designed to answer four questions:

(1) **Calibration accuracy.** Do EQC and SAC achieve a pre-specified target reliability ρ^* across a broad set of data-generating processes?
(2) **Robustness.** Is calibration performance stable under non-normal latent distributions and empirically realistic item pools?
(3) **Algorithmic agreement.** When both methods target the same estimand (average-information reliability $\tilde{\rho}$), do they yield similar calibrated scales c^*?

(4) **Reliability estimand choice.** How does the required calibration scale differ when SAC targets the MSEM-based reliability $\bar{w}$ versus the average-information reliability $\tilde{\rho}$, and do empirical results align with the Jensen's-inequality relationship $\tilde{\rho} \geq \bar{w}$ (Section 2.2)?

Experimental design and manipulated factors. We conducted a fully crossed factorial Monte Carlo study varying the following structural factors:

- **Latent distribution shape (G):** Normal, bimodal, positively skewed, and heavy-tailed (standardized to $\sigma_\theta^2 = 1$).
- **IRT model:** Rasch ($\lambda_{i,0} \equiv 1$) and 2PL (heterogeneous $\lambda_{i,0}$).
- **Item source:** Empirical item pools drawn from the Item Response Warehouse (IRW) versus parametric item pools.
- **Test length (I):** 15, 30, and 60 items.
- **Sample size for generated datasets (N):** 100, 200, 500, 1,000, and 2,000 persons.

For parametric item pools, item difficulties were generated as $\beta_i \sim \mathcal{N}(0, 1)$. Under the 2PL model, baseline discriminations were generated as $\log(\lambda_{i,0}) \sim \mathcal{N}(0, 0.3^2)$ with modest negative dependence between β_i and $\lambda_{i,0}$ induced via a Gaussian copula (target correlation $\rho = -0.3$) (Nelsen, 2006). Under the Rasch model, $\lambda_{i,0} \equiv 1$ by definition. For IRW pools, $(\beta_i, \lambda_{i,0})$ were sampled from the empirical joint distribution used by our IRW generator module. Full details of item parameter generation are provided in Appendix D.

The non-normal latent distributions were parameterized to represent common departures from normality: a symmetric bimodal mixture (mode separation $\delta = 0.8$), a positively skewed distribution (shape parameter $k = 4$), and a heavy-tailed distribution (Student-t with df $= 5$). Detailed specifications of these distributions are given in Appendix C.

Target reliability levels were adapted by test length to ensure feasibility (Section 2.4): $\rho^* \in \{0.30, 0.40, 0.50, 0.60\}$ for $I = 15$; $\rho^* \in \{0.40, 0.50, 0.60, 0.70\}$ for $I = 30$; and $\rho^* \in \{0.50, 0.60, 0.70, 0.80\}$ for $I = 60$. This scheme reflects the practical upper bounds on achievable reliability for short tests implied by the information-based limits in Section 2.4. In total, the design yields $4 \times 2 \times 2 \times 3 \times 5 \times 4 = 960$ conditions.

Calibration configuration and evaluation. EQC was implemented with a fixed quadrature of size $M = 20{,}000$ and deterministic root-finding over $c \in [0.1, 10]$.

Following diagnostic results indicating that the MSEM-based objective can be non-monotone in c for some item configurations, EQC was evaluated only for $\tilde{\rho}$.

SAC used 1,000 stochastic approximation iterations with 1,000 Monte Carlo draws per iteration (with Polyak–Ruppert averaging after burn-in) and was warm-started at the EQC solution to reduce burn-in. We ran two SAC variants: one targeting $\tilde{\rho}$ and one targeting $\bar{w}$.

Calibration accuracy was summarized by the deviation

$$\Delta \;=\; \rho_{\text{achieved}} - \rho^{*}, \tag{4.1}$$

where ρ_{achieved} denotes the reliability implied by the calibrated configuration, computed via Monte Carlo integration. For EQC, ρ_{achieved} corresponds to the empirical reliability function evaluated at the calibrated root; thus EQC deviations primarily reflect numerical root-finding tolerance and finite-precision arithmetic. For SAC, ρ_{achieved} was recomputed using an independent Monte Carlo evaluation sample, so deviations reflect stochastic approximation error and Monte Carlo variability.

Finally, to separate calibration error from finite-sample variability, we generated $K = 2{,}000$ replicated response datasets per condition using the EQC-calibrated configuration and each of the five sample sizes N. These replications quantify how much the *empirically realized* reliability varies across datasets even when the population-level design target is held fixed; replication-focused results are summarized in Appendix E.

4.2. **Results. Overall calibration accuracy.** Table 2 summarizes calibration accuracy across all 960 conditions. EQC achieved effectively perfect calibration of $\tilde{\rho}$ (MAE $\approx 10^{-5}$; 100% of conditions within even the strictest tolerance thresholds). SAC achieved calibration that was accurate on average but noisier: both SAC variants exhibited near-zero mean deviation (unbiasedness) with MAE ≈ 0.015 and SD ≈ 0.024.

The qualitative pattern in Table 2 is clarified by Figure 1, which plots achieved reliability against the target level across all conditions. EQC points lie exactly on the identity line. This is expected from the fixed-quadrature construction: conditional on the quadrature sample, EQC defines a deterministic empirical reliability function $\hat{\rho}_M(c)$ and finds c^* such that $\hat{\rho}_M(c^*) = \rho^*$. The reported achieved reliability is then $\rho_{\text{achieved}} = \hat{\rho}_M(c^*)$, which equals ρ^* up to numerical tolerance.

By contrast, SAC exhibits symmetric scatter around the identity line (Figure 1), reflecting stochastic approximation error and Monte Carlo variability. Importantly,

TABLE 2. Calibration Accuracy: Deviation from Target Reliability by Algorithm

Algorithm	Cond.	Mean Δ	SD	MAE	Max $\lvert\Delta\rvert$	$< .01$	$< .02$	$< .05$
EQC	960	−0.00000	0.00001	0.00001	0.0000	100.0%	100.0%	100.0%
SAC ($\tilde{\rho}$)	960	−0.00006	0.02396	0.01503	0.1693	53.3%	73.0%	94.4%
SAC ($\bar{w}$)	960	−0.00062	0.02437	0.01582	0.1792	50.7%	71.7%	94.9%

Note. Cond. = number of design conditions; Δ = deviation (achieved − target reliability); MAE = Mean Absolute Error. SAC ($\tilde{\rho}$) targets average-information reliability; SAC ($\bar{w}$) targets MSEM-based reliability. Percentages indicate the proportion of conditions with absolute deviation below each threshold.

the mean trend for SAC coincides with the identity line, confirming that SAC is approximately unbiased in achieving ρ^* at the design level. The remaining dispersion is expected under Robbins–Monro theory because (i) the SAC iterates converge at $O_p(n^{-1/2})$ and (ii) post-calibration evaluation uses independent Monte Carlo draws.

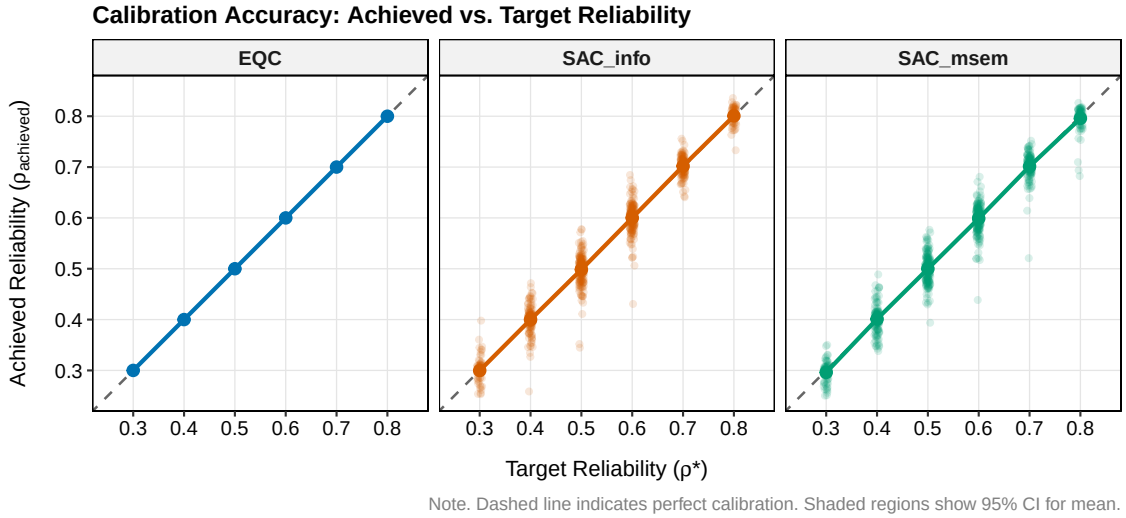

FIGURE 1. Calibration Accuracy: Achieved vs. Target Reliability

Note. The dashed diagonal line indicates perfect calibration ($\rho_{\text{achieved}} = \rho^*$). Semi-transparent points represent individual conditions. Solid lines show the condition-averaged achieved reliability for each algorithm.

Performance by target level. While Table 2 aggregates across targets, Table 3 shows achieved reliability by target level. EQC tracks each target essentially exactly (SD $\approx 10^{-5}$). SAC exhibits small mean discrepancies that remain close to the target at all levels, with slightly larger departures at the edges of the feasible range (e.g., $\rho^* = 0.30$ or 0.80). This edge behavior is consistent with the feasibility logic in

Section 2.4: near the lower and upper reliability bounds, the mapping between c and ρ becomes more sensitive to Monte Carlo error, and (for $\bar{w}$) the harmonic-mean structure of MSEM $= \mathbb{E}[1/\mathcal{J}(\theta)]$ increases sensitivity to low-information regions.

TABLE 3. Achieved Reliability by Target Level and Algorithm

ρ^*	Cond.	EQC Mean	EQC SD	SAC ($\tilde{\rho}$) Mean	SAC ($\bar{w}$) Mean
0.30	80	0.3000	0.00001	0.2997	0.2961
0.40	160	0.4000	0.00002	0.3999	0.4011
0.50	240	0.5000	0.00001	0.4984	0.4999
0.60	240	0.6000	0.00001	0.6000	0.5987
0.70	160	0.7000	0.00001	0.7020	0.7014
0.80	80	0.8000	0.00001	0.8007	0.7957

Note. ρ^* = target reliability; Cond. = number of design conditions at each target level. Adaptive targets were restricted to plausible ranges by test length: $I = 15$ ($\rho^* \in \{0.30, 0.40, 0.50, 0.60\}$), $I = 30$ ($\rho^* \in \{0.40, 0.50, 0.60, 0.70\}$), and $I = 60$ ($\rho^* \in \{0.50, 0.60, 0.70, 0.80\}$).

Robustness across latent distribution shapes. Figure 2 shows the distribution of deviations Δ across latent distribution shapes. EQC deviations are effectively zero across all shapes. For SAC, deviations remain centered near zero under all latent distributions, with broadly similar dispersion. Heavy-tailed and skewed distributions exhibit slightly more extreme outliers, which is expected because these distributions allocate more probability mass to trait regions where the test information function is low and more variable. Nevertheless, the median deviations remain close to zero, indicating that the calibration procedure remains stable even under substantial departures from normality.

Calibration accuracy by IRT model and item source. Figure 3 decomposes calibration deviations by IRT model and item source. The most notable pattern is that SAC variability is largest for 2PL conditions using IRW item pools. Empirical pools lead to irregular difficulties and heterogeneous discrimination structures, which induce greater variability in the test information function and consequently larger Monte Carlo variability in both $\tilde{\rho}$ and $\bar{w}$. By contrast, Rasch conditions—especially under parametric item generation—exhibit the tightest SAC deviations, consistent with the relative homogeneity of information when discriminations are fixed at unity.

Algorithm agreement: EQC vs. SAC discrimination scale. When targeting the same estimand ($\tilde{\rho}$), EQC and SAC should yield similar calibrated scales c^* (Section 3). Figure 4 compares the calibrated scale from EQC to the scale from SAC ($\tilde{\rho}$). Points cluster near the identity line, indicating strong agreement in the solution

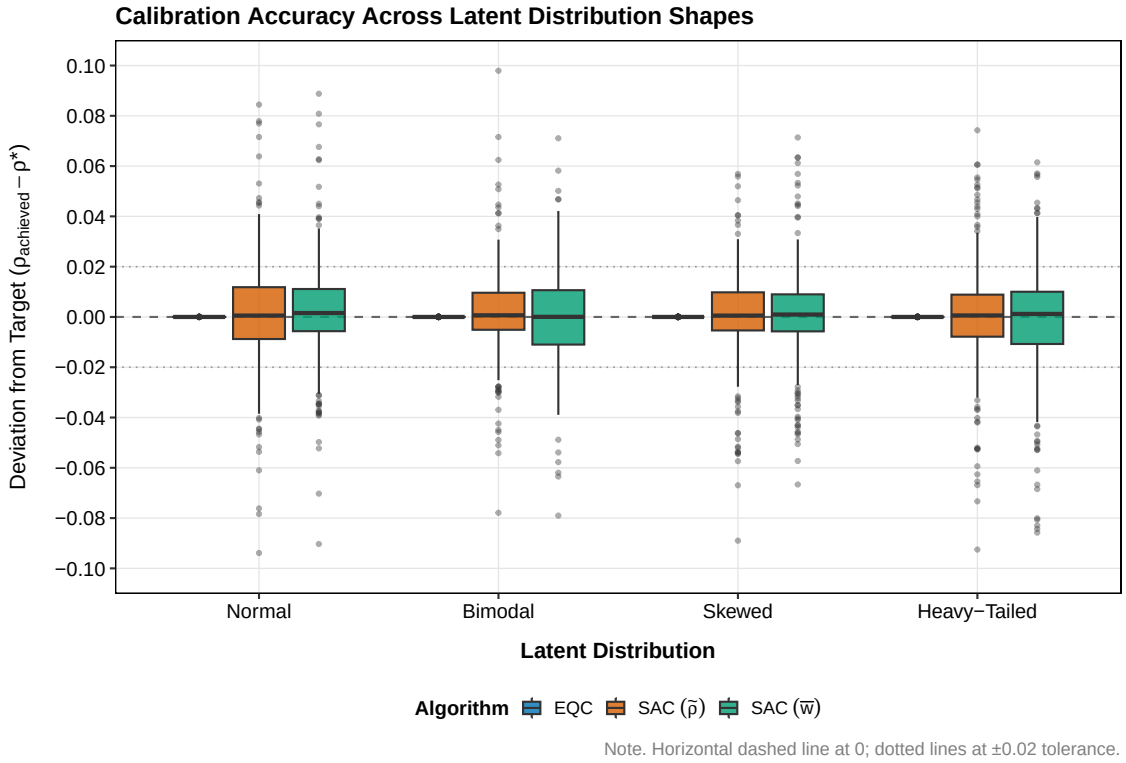

FIGURE 2. Calibration Accuracy Across Latent Distribution Shapes

Note. The outcome is the deviation from target reliability ($\Delta = \rho_{\text{achieved}} - \rho^*$). The horizontal dashed line is $\Delta = 0$; dotted lines denote a ± 0.02 tolerance band.

to the inverse design problem. The remaining dispersion is attributable to stochastic approximation variability and to the fact that EQC conditions on a fixed quadrature realization while SAC uses fresh Monte Carlo draws.

Jensen's inequality: SAC ($\tilde{\rho}$) vs. SAC ($\bar{w}$). Section 2.2 established that $\tilde{\rho}(c) \geq \bar{w}(c)$ by Jensen's inequality, because $\bar{w}$ depends on $\mathbb{E}[1/\mathcal{J}(\theta)]$ (a harmonic-mean structure) whereas $\tilde{\rho}$ depends on $\mathbb{E}[\mathcal{J}(\theta)]$ (an arithmetic-mean structure). Consequently, achieving the same target reliability should require a weakly larger scale when targeting $\bar{w}$ than when targeting $\tilde{\rho}$. Formally, if $c^*_{\tilde{\rho}}$ solves $\tilde{\rho}(c) = \rho^*$, then $\bar{w}(c^*_{\tilde{\rho}}) \leq \rho^*$, implying that the solution $c^*_{\bar{w}}$ satisfying $\bar{w}(c) = \rho^*$ must obey $c^*_{\bar{w}} \geq c^*_{\tilde{\rho}}$.

Figure 5 confirms this implication empirically: the SAC ($\bar{w}$) scales lie uniformly above the SAC ($\tilde{\rho}$) scales. The practical magnitude of this gap depends on the "Jensen gap" (how variable $\mathcal{J}(\theta)$ is across the population). When the test information function is nearly flat, $\tilde{\rho} \approx \bar{w}$ and the two scales nearly coincide; when information varies strongly across θ, the gap widens and SAC ($\bar{w}$) requires a meaningfully larger c^*.

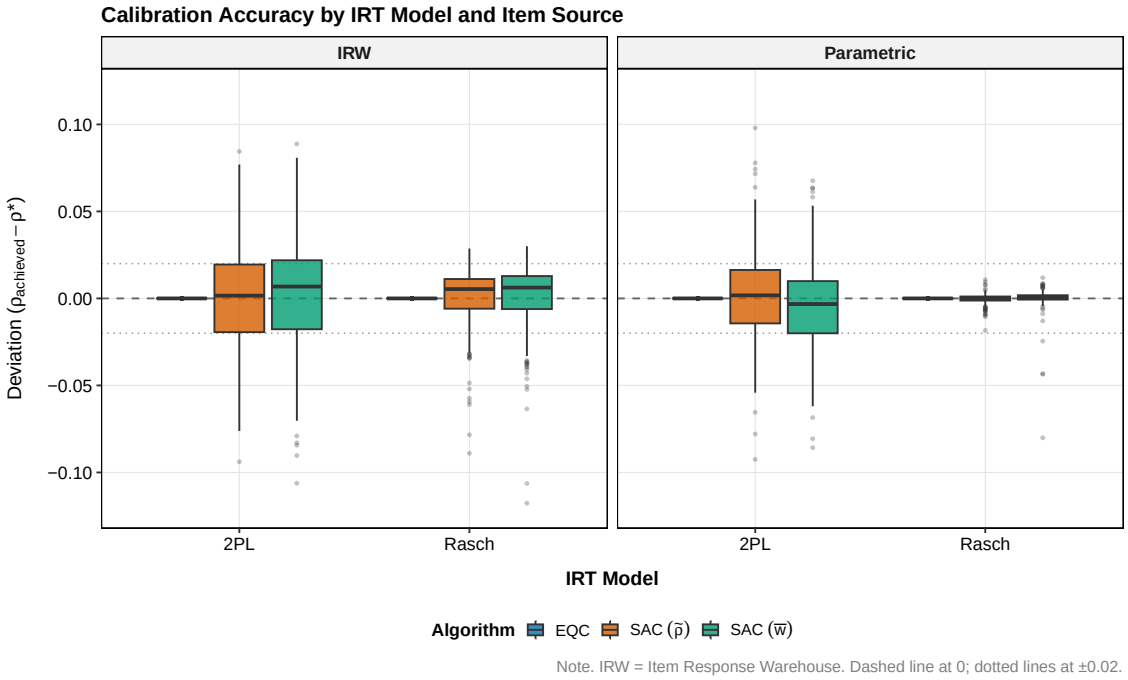

FIGURE 3. Calibration Accuracy by IRT Model and Item Source

Note. The outcome is the deviation from target reliability ($\Delta = \rho_{\text{achieved}} - \rho^*$). The horizontal dashed line is $\Delta = 0$; dotted lines denote a ± 0.02 tolerance band. IRW = Item Response Warehouse.

In particular, heavy-tailed latent distributions exhibit a somewhat larger gap, because they allocate more probability mass to extreme θ values where test information is typically lower.

Summary. The validation study yields three primary conclusions. First, EQC solves the inverse design problem for $\tilde{\rho}$ to numerical precision. Across all 960 conditions, EQC deviations are essentially zero, reflecting deterministic root-finding on a fixed empirical quadrature.

Second, SAC provides an accurate, unbiased calibration method with predictable stochastic variability, and uniquely enables direct calibration to $\bar{w}$. Although SAC is noisier than EQC, its deviations remain centered at zero, and a large majority of conditions fall within practically useful tolerance bands (e.g., ± 0.02 or ± 0.05).

Third, the empirical results confirm the theoretical implications of estimand choice. EQC and SAC agree closely when targeting $\tilde{\rho}$, and the calibrated scale required to target $\bar{w}$ is systematically larger than that required to target $\tilde{\rho}$—a direct consequence of Jensen's inequality. Together, these findings validate both the algorithmic toolkit

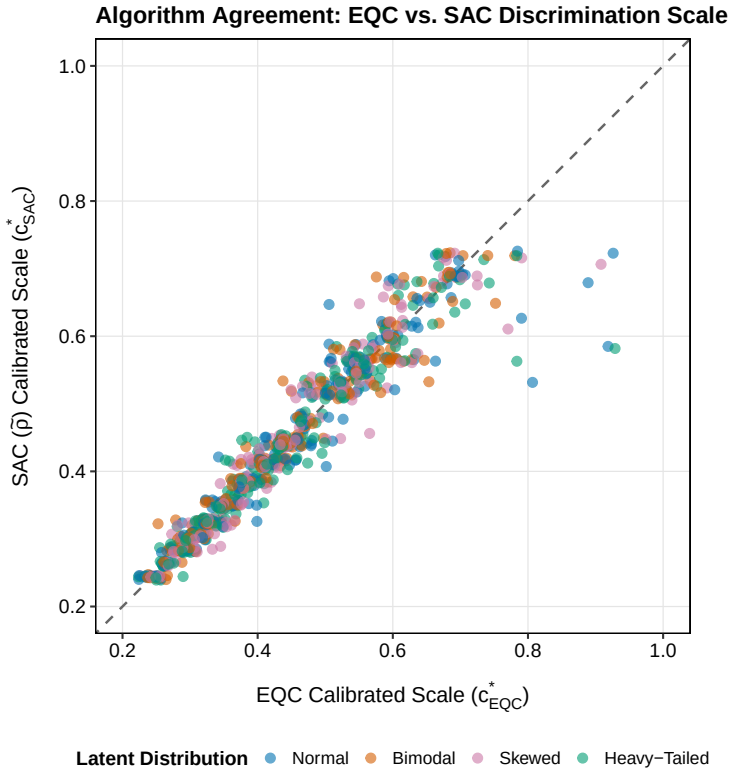

FIGURE 4. Algorithm Agreement: EQC vs. SAC Discrimination Scale

Note. The dashed diagonal line indicates perfect agreement in the calibrated scale ($c^*_{\text{EQC}} = c^*_{\text{SAC}}$). Both algorithms target average-information reliability $\tilde{\rho}$. Point colors indicate the latent distribution shape.

(Section 3) and the reliability framework (Section 2), supporting reliability-targeted simulation as a practical and theoretically grounded design strategy.

5. Discussion and Conclusion

Reliability is the central signal-to-noise quantity in measurement-based simulation: it determines how much information about the latent trait is present in generated responses and, consequently, how difficult the downstream estimation problem truly is. Yet IRT simulation studies have rarely treated marginal reliability as an explicit design factor, despite its direct interpretability and its role in determining ecological realism and cross-condition comparability. The framework developed in this paper closes this gap by turning marginal reliability from an incidental by-product of item and population generation into an explicit input—one that can be targeted, varied, and reported with the same status as N, I, and other design factors.

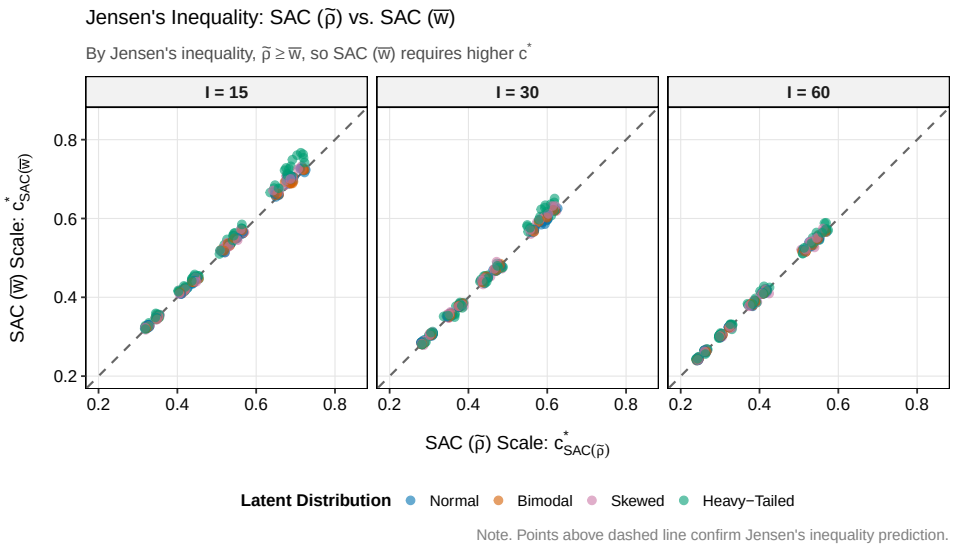

FIGURE 5. Jensen's Inequality: SAC ($\tilde{\rho}$) vs. SAC ($\bar{w}$)

Note. The dashed diagonal line indicates equality of scales ($c^*_{\mathrm{SAC}(\tilde{\rho})} = c^*_{\mathrm{SAC}(\bar{w})}$). Points above the line indicate that targeting $\bar{w}$ requires a larger calibration scale, consistent with Jensen's inequality ($\tilde{\rho} \geq \bar{w}$).

5.1. **Summary of Contributions.** This paper makes three contributions.

First, we formalize *reliability-targeted IRT simulation* as an inverse design problem. Under a specified structural configuration—latent distribution G, item-parameter distributions, and test length—we seek a global discrimination scaling factor c^* such that the induced population reliability equals a target level ρ^*. The key idea is to separate *structure* from *scale*: realistic item and population features are generated first, and then overall measurement strength is tuned by a single multiplier applied to discrimination. Concretely, if $\lambda_{i,0}$ denotes a baseline discrimination from the chosen item generator, the calibrated discrimination is $\lambda_i^* = c^* \lambda_{i,0}$. Under mild regularity conditions, the reliability function is strictly increasing in c, which yields existence and uniqueness of the solution c^* for any feasible target. This theoretical foundation clarifies that reliability is not a fixed attribute of a model class but a functional of the full data-generating process—and that it can be controlled without sacrificing realism.

Second, we introduce two complementary calibration algorithms for solving the inverse design problem. Empirical Quadrature Calibration (EQC) is designed for routine use: conditional on a fixed quadrature sample, it constructs a deterministic

empirical approximation to the population reliability function and uses numerical root finding to obtain c^* rapidly and stably. Stochastic Approximation Calibration (SAC) provides a more general alternative based on Robbins–Monro updates and Polyak–Ruppert averaging. SAC is particularly useful when deterministic quadrature is inconvenient, when the user wants an independent stochastic check of an EQC-based design, or when targeting reliability estimands that are most naturally evaluated via Monte Carlo.

Third, we validate the framework and algorithms in a comprehensive Monte Carlo study varying latent distribution shape, IRT model (Rasch vs. 2PL), item source (empirical vs. parametric), test length, and target reliability. The results show that EQC attains essentially exact calibration in the targeted estimand, while SAC remains approximately unbiased with quantifiable Monte Carlo variability—accuracy levels that are sufficient for most practical simulation research. The validation also clarifies a key interpretive point: average-information and MSEM-based marginal reliabilities are not numerically interchangeable. Because Jensen's inequality implies a systematic gap between the arithmetic-mean and harmonic-mean information functionals, designs that target average-information reliability and designs that target MSEM-based reliability will generally require different calibrated scales, even under identical structural settings. This is not a flaw of the calibration; it reflects a substantive distinction in what "reliability" is being held constant.

5.2. **Practical Recommendations for Simulation Researchers.** Several pragmatic guidelines follow directly from the theoretical results and validation patterns.

Treat marginal reliability as a primary design factor. In many simulation settings, the most consequential difference between "easy" and "hard" datasets is not sample size or estimator choice but the amount of measurement information present in the responses. Reliability should therefore be (i) explicitly targeted during design, (ii) varied as an experimental factor when studying robustness, and (iii) reported alongside other design features such as N, I, latent distribution shape, and item pool characteristics. This practice improves ecological validity and reduces the risk that apparent method differences are artifacts of uncontrolled information regimes.

Choose the reliability estimand to match the substantive goal. This paper distinguishes average-information reliability (denoted $\tilde{\rho}$) and MSEM-based marginal reliability (denoted $\bar{w}$). Average-information reliability aligns with an "expected information" description of measurement quality and is computationally convenient; $\bar{w}$ corresponds directly to a variance-decomposition definition based on the expected

conditional error variance. Because these estimands are not numerically interchangeable (and satisfy $\tilde{\rho} \geq \bar{w}$ under the conditions described in Section 2.2), simulation designs should state which estimand is being targeted. As a practical rule, targeting $\bar{w}$ is preferable when the evaluation criterion is tightly tied to conditional error variance (or when a conservative target is desired), whereas targeting $\tilde{\rho}$ is sensible when the goal is to align conditions by an information-based summary of overall test precision—provided the interpretation is made explicit.

Check feasibility early and interpret boundary behavior correctly. For short tests or restrictive item pools, some targets may be unattainable even with extreme scaling. Researchers should therefore compute (or at least approximate) achievable reliability bounds implied by their structural configuration and then either (i) adjust the target, (ii) increase test length, or (iii) widen the admissible scaling interval. When targets lie near feasibility limits, small changes in discrimination scaling can produce disproportionately large changes in reliability; such "edge targets" should be used intentionally rather than inadvertently.

Use EQC for routine calibration, and SAC for generality or independent validation. A practical workflow is to calibrate with EQC to obtain c^* quickly and deterministically, and then (optionally) run SAC initialized at the EQC solution to provide an independent stochastic check or to target $\bar{w}$ directly. Warm-starting SAC from EQC reduces burn-in and improves efficiency. If strict adherence to narrow tolerance bands is required, EQC is strongly preferred; if modest stochastic variability is acceptable—or if the target estimand is most naturally handled via Monte Carlo—SAC provides a flexible and theoretically grounded alternative.

Finally, even when population-level reliability is tightly controlled, the *realized* reliability in finite samples will vary across replications. For studies aiming to mimic finite-sample behavior closely, it is useful to distinguish (a) the design-level target (controlled by calibration) from (b) replication-level variability induced by finite N and finite test length. Reporting both quantities helps readers separate calibration performance from irreducible variability in realized precision across simulated datasets.

5.3. **Limitations and Future Directions.** Several limitations define the current scope and motivate future work.

Scope of measurement models. The present study focuses on dichotomous Rasch and 2PL models. Extending reliability-targeted calibration to additional models—such as the 3PL, polytomous models (e.g., graded response and partial credit), and

multidimensional IRT—is a natural next step. These extensions are conceptually straightforward because the proposed algorithms require only (i) model-specific routines to compute test information (or its analogue) and (ii) evaluation of the corresponding marginal reliability functional $\rho(c)$ under a global scaling of the relevant slope/discrimination parameters. For polytomous models, test information generalizes as a sum of category-response information contributions that retain monotonic scaling in discrimination, so the EQC root-finding logic and SAC updates carry over with minimal modification. For multidimensional models, discrimination becomes a vector and information is typically matrix-valued; one can apply a global scalar multiplier to discrimination-vector magnitudes (preserving directional structure) and calibrate a scalar summary of reliability (e.g., trace- or determinant-based summaries), or consider dimension-specific scaling when dimension-wise targets are desired. For the 3PL, the presence of a guessing parameter may compress achievable reliability ranges, but global discrimination scaling remains a viable calibration lever, and the inverse design formulation remains applicable.

What is controlled: global reliability, not the shape of precision. A single scaling factor controls the overall magnitude of measurement information but does not change where along the latent continuum the test is most informative. If a simulation study requires targeted precision at specific trait levels (e.g., floor/ceiling designs, adaptive testing, or differential precision across subpopulations), reliability targeting should be combined with additional structural manipulations such as the difficulty distribution, test targeting, or adaptive item selection rules. In this sense, the proposed framework is best viewed as a modular "reliability layer" that can be added after other structural design choices are set.

Finite-sample realism and estimator-specific reporting. The primary estimands targeted in this paper are population-level quantities. In applied reporting, however, reliability is often expressed through estimator-dependent measures (e.g., WLE vs. EAP reliability), and finite-sample estimation introduces additional variability beyond the design target. Future work should further clarify how to map design-level targets to commonly reported estimator-based reliabilities under finite-sample estimation, and how robust such mappings are under model misspecification or atypical response behavior.

5.4. **Conclusion.** Reliability-targeted simulation is both theoretically well founded and practically achievable. By treating marginal reliability as an explicit control

variable—analogous to the role of ICC in multilevel simulation—researchers can design IRT simulation studies that are more interpretable, more ecologically realistic, and more reproducible. The proposed EQC and SAC algorithms, together with the `IRTsimrel` implementation, lower the barrier to adopting this practice and encourage routine reporting and manipulation of reliability as a first-class design parameter. We anticipate that making "Target Reliability: ρ^*" a standard line in simulation protocols will improve the comparability and cumulative value of psychometric simulation evidence.

References

Adams, R. J. (2005). Reliability as a measurement design effect. *Studies in Educational Evaluation*, *31*(2-3), 162–172. https://doi.org/10.1016/j.stueduc.2005.05.008 [6].

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd). CRC Press. [6].

Bambirra Gonçalves, F., da Costa Campos Dias, B., & Machado Soares, T. (2018). Bayesian item response model: A generalized approach for the abilities' distribution using mixtures. *Journal of Statistical Computation and Simulation*, *88*(5), 967–981. https://doi.org/10.1080/00949655.2017.1413650 [2].

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395–479). Addison-Wesley. [5].

Brennan, R. L., & Kane, M. T. (1977). Signal/noise ratios for domain-referenced tests. *Psychometrika*, *42*(4), 609–625. https://doi.org/10.1007/BF02295983 [2].

Brent, R. P. (1971). An algorithm with guaranteed convergence for finding a zero of a function. *The Computer Journal*, *14*(4), 422–425. https://doi.org/10.1093/comjnl/14.4.422 [11].

Can, S., van de Schoot, R., & Hox, J. (2015). Collinear latent variables in multilevel confirmatory factor analysis: A comparison of maximum likelihood and Bayesian estimations. *Educational and Psychological Measurement*, *75*(3), 406–427. https://doi.org/10.1177/0013164414547959 [3].

Cheng, S., & Meng, X. (2025). Estimating IRT models under Gaussian mixture modeling of latent traits: An application of MSAEM algorithm. *Multivariate Behavioral Research*, 1–18. https://doi.org/10.1080/00273171.2025.2512345 [2].

Cho, S.-J., Preacher, K. J., & Bottge, B. A. (2015). Detecting intervention effects in a cluster-randomized design using multilevel structural equation modeling for binary responses. *Applied Psychological Measurement*, *39*(8), 627–642. https://doi.org/10.1177/0146621615591094 [3].

Conoyer, S. J., Therrien, W. J., & White, K. K. (2022). Meta-analysis of validity and review of alternate form reliability and slope for curriculum-based measurement in science and social studies. *Assessment for Effective Intervention*, *47*(2), 101–111. https://doi.org/10.1177/1534508420978457 [3].

Cronbach, L. J., & Gleser, G. C. (1964). The signal/noise ratio in the comparison of reliability coefficients. *Educational and Psychological Measurement*, *24*(3), 467–480. https://doi.org/10.1177/001316446402400303 [2].

de Ayala, R. J. (2022). *The theory and practice of item response theory* (2nd). Guilford Press. [6].

Domingue, B. W., Braginsky, M., Caffrey-Maffei, L., Gilbert, J. B., Kanopka, K., Kapoor, R., Lee, H., Liu, Y., Nadela, S., Pan, G., Zhang, L., Zhang, S., & Frank, M. C. (2025). An introduction to the Item Response Warehouse (IRW): A resource for enhancing data usage in psychometrics. *Behavior Research Methods*, *57*(10), 276. https://doi.org/10.3758/s13428-025-02796-y [4, 49].

Doran, H. C. (2005). The information function for the one-parameter logistic model: Is it reliability? *Educational and Psychological Measurement*, *65*(5), 665–675. https://doi.org/10.1177/0013164404272500 [7].

Fischer, G. H., & Molenaar, I. W. (Eds.). (1995). *Rasch models: Foundations, recent developments, and applications*. Springer. https://doi.org/10.1007/978-1-4612-4230-7 [6].

Frisbie, D. A. (1988). Reliability of scores from teacher-made tests. *Educational Measurement: Issues and Practice*, *7*(1), 25–35. https://doi.org/10.1111/j.1745-3992.1988.tb00422.x [3].

Gadat, S., & Panloup, F. (2023). Optimal non-asymptotic analysis of the Ruppert–Polyak averaging stochastic algorithm. *Stochastic Processes and Their Applications*, *156*, 312–348. https://doi.org/10.1016/j.spa.2022.11.012 [12].

Gilholm, P., Mengersen, K., & Thompson, H. (2021). Bayesian hierarchical multidimensional item response modeling of small sample, sparse data for personalized developmental surveillance. *Educational and Psychological Measurement*, *81*(5), 936–956. https://doi.org/10.1177/0013164420987582 [3].

Guastadisegni, L., Cagnone, S., Moustaki, I., & Vasdekis, V. (2025). The generalized Hausman test for detecting non-normality in the latent variable distribution of the two-parameter IRT model. *British Journal of Mathematical and Statistical Psychology*, *78*(3), 734–756. https://doi.org/10.1111/bmsp.12379 [2].

Guyatt, G. H., Kirshner, B., & Jaeschke, R. (1992). Measuring health status: What are the necessary measurement properties? *Journal of Clinical Epidemiology*, *45*(12), 1341–1345. https://doi.org/10.1016/0895-4356(92)90194-R [2].

Hsu, H.-Y., Lin, J.-H., Kwok, O.-M., Acosta, S., & Willson, V. (2017). The impact of intraclass correlation on the effectiveness of level-specific fit indices in multilevel structural equation modeling: A Monte Carlo study. *Educational and Psychological Measurement*, *77*(1), 5–31. https://doi.org/10.1177/0013164416642823 [3].

Lee, J., Che, J., Rabe-Hesketh, S., Feller, A., & Miratrix, L. (2025). Improving the estimation of site-specific effects and their distribution in multisite trials. *Journal of Educational and Behavioral Statistics*, *50*(5), 731–764. https://doi.org/10.3102/10769986241254286 [3, 6].

Lee, J., & Wind, S. A. (2025). Targeting toward inferential goals in Bayesian Rasch models for estimating person-specific latent traits [OSF Preprint]. https://osf.io/preprints/osf/qrw4n_v1 [3].

Lüdtke, O., Robitzsch, A., & Grund, S. (2017). Multiple imputation of missing data in multilevel designs: A comparison of different strategies. *Psychological Methods*, *22*(1), 141–165. https://doi.org/10.1037/met0000096 [3].

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, *1*(3), 86–92. https://doi.org/10.1027/1614-2241.1.3.86 [3].

McNeish, D. M., & Stapleton, L. M. (2016). The effect of small sample size on two-level model estimates: A review and illustration. *Educational Psychology Review*, *28*(2), 295–314. https://doi.org/10.1007/s10648-014-9287-x [3].

Monroe, S., & Cai, L. (2014). Estimation of a Ramsay-curve item response theory model by the Metropolis–Hastings Robbins–Monro algorithm. *Educational and Psychological Measurement*, *74*(2), 343–369. https://doi.org/10.1177/0013164413499344 [2].

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11), 2074–2102. https://doi.org/10.1002/sim.8086 [2].

Nelsen, R. B. (2006). *An introduction to copulas* (2nd). Springer. https://doi.org/10.1007/0-387-28678-0 [14].

Paddock, S. M., Ridgeway, G., Lin, R., & Louis, T. A. (2006). Flexible distributions for triple-goal estimates in two-stage hierarchical models. *Computational Statistics & Data Analysis*, *50*(11), 3243–3262. https://doi.org/10.1016/j.csda.2005.05.008 [3].

Paganin, S., Paciorek, C. J., Wehrhahn, C., Rodríguez, A., Rabe-Hesketh, S., & de Valpine, P. (2022). Computational strategies and estimation performance with Bayesian semiparametric item response theory models. *Journal of Educational and Behavioral Statistics*, *48*(2), 147–188. https://doi.org/10.3102/10769986221136105 [2].

Polyak, B. T., & Juditsky, A. B. (1992). Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, *30*(4), 838–855. https://doi.org/10.1137/0330046 [12].

Rabe-Hesketh, S., & Skrondal, A. (2022). *Multilevel and longitudinal modeling using Stata* (4th). Stata Press. [6].

Ramsay, J. O. (2016). Functional approaches to modeling response data. In W. J. van der Linden (Ed.), *Handbook of item response theory, volume one: Models* (pp. 365–378). CRC Press. https://doi.org/10.1201/9781315374512 [3].

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests* [Original work published 1960]. University of Chicago Press. [5].

Robbins, H., & Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, *22*(3), 400–407. http://www.jstor.org/stable/2236626 [4, 12].

Rouder, J. N., & Mehrvarz, M. (2024). Hierarchical-model insights for planning and interpreting individual-difference studies of cognitive abilities. *Current Directions in Psychological Science*, *33*(2), 128–135. https://doi.org/10.1177/09637214231220923 [2].

Soland, J., Kuhfeld, M., & Edwards, K. (2024). How survey scoring decisions can influence your study's results: A trip through the IRT looking glass. *Psychological Methods*, *29*(5), 1003–1024. https://doi.org/10.1037/met0000506 [3].

Sweeney, S. M., Sinharay, S., Johnson, M. S., & Steinhauer, E. W. (2022). An investigation of the nature and consequence of the relationship between IRT difficulty and discrimination. *Educational Measurement: Issues and Practice*, *41*(4), 50–67. https://doi.org/10.1111/emip.12522 [49, 50].

Thissen, D., & Wainer, H. (2001). *Test scoring*. Lawrence Erlbaum. [6].

Toulis, P., Horel, T., & Airoldi, E. M. (2021). The proximal Robbins–Monro method. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *83*(1), 188–212. https://doi.org/10.1111/rssb.12405 [12].

van der Linden, W. J. (2005). *Linear models for optimal test design*. Springer. https://doi.org/10.1007/0-387-29054-0 [7].

Wang, C., Su, S., & Weiss, D. J. (2018). Robustness of parameter estimation to assumptions of normality in the multidimensional graded response model. *Multivariate Behavioral Research*, *53*(3), 403–418. https://doi.org/10.1080/00273171.2018.1455572 [2].

Woods, C. M., & Thissen, D. (2006). Item response theory with estimation of the latent population distribution using spline-based densities. *Psychometrika*, *71*(2), 281–301. https://doi.org/10.1007/s11336-004-1175-8 [2].

Yang, J. S., Hansen, M., & Cai, L. (2012). Characterizing sources of uncertainty in item response theory scale scores. *Educational and Psychological Measurement*, *72*(2), 264–290. https://doi.org/10.1177/0013164411410056 [6].

Zhang, L., Liu, Y., Molenaar, D., Gilbert, J., Kanopka, K., & Domingue, B. W. (2025). Realistic simulation of item difficulties [OSF Preprint]. https://doi.org/10.31234/osf.io/jbhxy_v4 [4, 49].

Zumbo, B. D. (2025). Reliability as projection in operator-theoretic test theory: Conditional expectation, Hilbert space geometry, and implications for psychometric practice. *Educational and Psychological Measurement*. https://doi.org/10.1177/00131644251389891 [3].

Online Appendix

December 19, 2025

Contents

Appendix A. Mathematical Proofs and Derivations

This appendix provides formal derivations supporting the theoretical claims in Sections 2 and 3 of the main text. In particular, we (i) derive closed-form derivatives of item and test information under global discrimination scaling, (ii) formalize the monotonicity logic that makes the inverse design problem well-posed on a practical calibration interval, (iii) prove the Jensen-inequality ordering between the two reliability estimands, and (iv) state standard large-sample guarantees for EQC and SAC. Achievable reliability bounds are treated separately in Appendix B.

A.1. Notation, Setup, and Standing Conditions.

A.1.1. *Measurement model and global discrimination scaling.* We work with the dichotomous 2PL logistic model (Equation (2.1)). For a fixed test form with item difficulties and baseline discriminations

$$\Psi = \{(\beta_i, \lambda_{i,0})\}_{i=1}^{I}, \qquad \lambda_{i,0} > 0,$$

global discrimination scaling (Equation (2.5)) is

$$\lambda_i(c) = c\,\lambda_{i,0}, \qquad c > 0,\ i = 1, \ldots, I. \tag{A.1}$$

Define the conditional success probability

$$\pi_i(\theta; c) = \Pr(Y_i = 1 \mid \theta; \beta_i, \lambda_i(c)) = \frac{\exp\{\lambda_i(c)(\theta - \beta_i)\}}{1 + \exp\{\lambda_i(c)(\theta - \beta_i)\}}. \tag{A.2}$$

Let $\theta \sim G$ with $\mathbb{E}[\theta] = 0$ and $\mathrm{Var}(\theta) = \sigma_\theta^2 \in (0, \infty)$, as assumed in Section 2.

A.1.2. *Test information and reliability functionals.* The item Fisher information (for θ) under the 2PL logistic model is

$$\mathcal{J}_i(\theta; c) = \lambda_i(c)^2\, \pi_i(\theta; c)\{1 - \pi_i(\theta; c)\}. \tag{A.3}$$

The test information (Equation (2.2)) is

$$\mathcal{J}(\theta; c) = \sum_{i=1}^{I} \mathcal{J}_i(\theta; c). \tag{A.4}$$

Two population-level reliability functionals (Section 2.2) are:

MSEM-based marginal reliability (Equation (2.3)).

$$\mathrm{MSEM}(c) = \mathbb{E}\left[\frac{1}{\mathcal{J}(\theta; c)}\right], \qquad \bar{w}(c) = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \mathrm{MSEM}(c)}. \tag{A.5}$$

Average-information reliability (Equation (2.4)).

$$\bar{\mathcal{J}}(c) = \mathbb{E}\big[\mathcal{J}(\theta; c)\big], \qquad \tilde{\rho}(c) = \frac{\sigma_\theta^2\, \bar{\mathcal{J}}(c)}{\sigma_\theta^2\, \bar{\mathcal{J}}(c) + 1}. \tag{A.6}$$

A.1.3. *Regularity conditions on a practical calibration interval.* Throughout, calibration is restricted to a compact interval $[c_L, c_U] \subset (0, \infty)$ as in Section 2.3. The substantive "well-posedness" statements in the main text require two types of conditions: (i) finiteness and interchange of differentiation and expectation, and (ii) monotonicity of the relevant derivative expectations on $[c_L, c_U]$.

Condition A.1 (Positivity and finiteness). For all $c \in [c_L, c_U]$,

$$\mathbb{E}\big[\mathcal{J}(\theta; c)\big] < \infty, \qquad \mathbb{E}\left[\frac{1}{\mathcal{J}(\theta; c)}\right] < \infty. \tag{A.7}$$

(Under the logistic model, $\mathcal{J}(\theta; c) > 0$ for all finite (θ, c), but $\mathbb{E}[1/\mathcal{J}]$ can diverge if G places non-negligible mass in regions where $\mathcal{J}$ is extremely small.)

Condition A.2 (Differentiation under the integral sign). For $c \in [c_L, c_U]$, $\mathcal{J}(\theta; c)$ is differentiable in c for a.e. θ, and there exist integrable envelopes $M_1(\theta), M_2(\theta)$ such that for all $c \in [c_L, c_U]$,

$$\left|\frac{\partial}{\partial c}\mathcal{J}(\theta; c)\right| \le M_1(\theta), \qquad \left|\frac{\partial}{\partial c}\left(\frac{1}{\mathcal{J}(\theta; c)}\right)\right| = \left|\frac{-\mathcal{J}'(\theta; c)}{\mathcal{J}(\theta; c)^2}\right| \le M_2(\theta), \tag{A.8}$$

with $\mathbb{E}[M_1(\theta)] < \infty$ and $\mathbb{E}[M_2(\theta)] < \infty$.

Condition A.3 (Monotonicity on the practical interval). On $[c_L, c_U]$,

$$\bar{\mathcal{J}}'(c) = \frac{d}{dc}\mathbb{E}\big[\mathcal{J}(\theta; c)\big] > 0, \qquad \mathbb{E}\left[\frac{\mathcal{J}'(\theta; c)}{\mathcal{J}(\theta; c)^2}\right] > 0. \tag{A.9}$$

This formalizes the "dense item grid / no information holes" regime discussed in Section 2.3: extremely large scaling can produce information spikes and gaps, which may cause non-monotonicity in MSEM-based reliability; restricting to a practical $[c_L, c_U]$ avoids this regime.

A.2. Derivatives of Information Under Global Discrimination Scaling.

A.2.1. *Logistic kernel representation.* Let $s(x) = \{1 + \exp(-x)\}^{-1}$ and define the logistic variance kernel

$$h(x) = s(x)\{1 - s(x)\} = \frac{e^x}{(1 + e^x)^2}. \tag{A.10}$$

With

$$x_i(\theta; c) = \lambda_i(c)(\theta - \beta_i) = c\,\lambda_{i,0}(\theta - \beta_i),$$

Equation (A.3) can be written as

$$\mathcal{J}_i(\theta; c) = \lambda_i(c)^2\, h\{x_i(\theta; c)\} = c^2\lambda_{i,0}^2\, h(c\lambda_{i,0}(\theta - \beta_i))\,. \tag{A.11}$$

A.2.2. *Derivative of item information and local (non-)monotonicity.*

Lemma A.1 (Derivative of $\mathcal{J}_i(\theta; c)$)**.** *For each fixed (θ, i), $\mathcal{J}_i(\theta; c)$ is differentiable in c, and*

$$\frac{\partial}{\partial c}\mathcal{J}_i(\theta; c) = c\,\lambda_{i,0}^2\, h\{x_i(\theta; c)\}\left[2 - x_i(\theta; c)\tanh\big(x_i(\theta; c)/2\big)\right]. \tag{A.12}$$

Proof. Differentiate Equation (A.11). Let $x = x_i(\theta; c)$. Then $\mathcal{J}_i = c^2\lambda_{i,0}^2 h(x)$ and $dx/dc = \lambda_{i,0}(\theta - \beta_i) = x/c$. Since $h'(x) = h(x)\{1 - 2s(x)\} = -h(x)\tanh(x/2)$,

$$\begin{aligned}\frac{\partial}{\partial c}\mathcal{J}_i &= 2c\lambda_{i,0}^2 h(x) + c^2\lambda_{i,0}^2 h'(x)\frac{x}{c}\\ &= c\lambda_{i,0}^2 h(x)\Big[2 + x\{1 - 2s(x)\}\Big]\\ &= c\lambda_{i,0}^2 h(x)\Big[2 - x\tanh(x/2)\Big].\end{aligned}$$

□

Remark 1 (Local non-monotonicity)**.** Define $\phi(x) = 2 - x\tanh(x/2)$. For $x > 0$,

$$\phi'(x) = -\tanh(x/2) - \frac{x}{2}\mathrm{sech}^2(x/2) < 0,$$

so ϕ is strictly decreasing on $(0, \infty)$ with $\phi(0) = 2$ and $\lim_{x\to\infty}\phi(x) = 2 - x$. Hence ϕ has a unique positive root $x_0 \approx 2.399$, and $\phi(x) > 0$ iff $|x| < x_0$. Therefore, for fixed (θ, i), $\mathcal{J}_i(\theta; c)$ increases in c when $|x_i(\theta; c)| < x_0$, but can decrease once $|x_i(\theta; c)|$ is sufficiently large. This mechanism explains why extreme scaling can generate information spikes near item difficulties and information gaps between them, motivating restriction to a practical calibration interval.

A.2.3. *Derivative of test information.* Summing Equation (A.12) over items yields

$$\mathcal{J}'(\theta; c) \equiv \frac{\partial}{\partial c}\mathcal{J}(\theta; c) = \sum_{i=1}^{I} c\,\lambda_{i,0}^2\, h\{x_i(\theta; c)\}\left[2 - x_i(\theta; c)\tanh\big(x_i(\theta; c)/2\big)\right]. \tag{A.13}$$

A.3. **Monotonicity of Reliability and Existence/Uniqueness of the Calibrated Scale.** This section formalizes the monotonicity claim in Section 2.3 and the resulting existence/uniqueness of the calibrated scale.

A.3.1. *Derivatives of $\tilde{\rho}(c)$ and $\bar{w}(c)$.*

Lemma A.2 (Derivatives of reliability functionals)**.** *Assume Conditions A.1–A.2. Then for* $c \in [c_L, c_U]$*,*

$$\tilde{\rho}'(c) = \frac{\sigma_\theta^2}{\{\sigma_\theta^2 \bar{\mathcal{J}}(c) + 1\}^2} \bar{\mathcal{J}}'(c), \qquad \bar{\mathcal{J}}'(c) = \mathbb{E}\big[\mathcal{J}'(\theta; c)\big], \tag{A.14}$$

and

$$\bar{w}'(c) = \frac{\sigma_\theta^2}{\{\sigma_\theta^2 + \mathrm{MSEM}(c)\}^2} \mathbb{E}\left[\frac{\mathcal{J}'(\theta; c)}{\mathcal{J}(\theta; c)^2}\right]. \tag{A.15}$$

Proof. For $\tilde{\rho}$, write $\tilde{\rho}(c) = r(\bar{\mathcal{J}}(c))$ with $r(u) = \sigma_\theta^2 u/(\sigma_\theta^2 u+1)$, so $r'(u) = \sigma_\theta^2/(\sigma_\theta^2 u+1)^2$. Condition A.2 permits $\bar{\mathcal{J}}'(c) = \frac{d}{dc}\mathbb{E}[\mathcal{J}] = \mathbb{E}[\mathcal{J}']$, giving Equation (A.14).

For $\bar{w}(c) = \sigma_\theta^2/(\sigma_\theta^2 + \mathrm{MSEM}(c))$, we have

$$\bar{w}'(c) = -\frac{\sigma_\theta^2}{\{\sigma_\theta^2 + \mathrm{MSEM}(c)\}^2} \mathrm{MSEM}'(c).$$

By Condition A.2 and $\mathrm{MSEM}(c) = \mathbb{E}[1/\mathcal{J}(\theta; c)]$,

$$\mathrm{MSEM}'(c) = \mathbb{E}\left[\frac{\partial}{\partial c}\left(\frac{1}{\mathcal{J}(\theta; c)}\right)\right] = \mathbb{E}\left[-\frac{\mathcal{J}'(\theta; c)}{\mathcal{J}(\theta; c)^2}\right],$$

which yields Equation (A.15). □

A.3.2. *Monotonicity and uniqueness on* $[c_L, c_U]$*.*

Proposition A.1 (Strict monotonicity on the practical interval)**.** *Under Conditions A.1–A.3, both* $\tilde{\rho}(c)$ *and* $\bar{w}(c)$ *are strictly increasing on* $[c_L, c_U]$*.*

Proof. By Lemma A.2, $\tilde{\rho}'(c)$ has the same sign as $\bar{\mathcal{J}}'(c)$, and $\bar{w}'(c)$ has the same sign as $\mathbb{E}[\mathcal{J}'/\mathcal{J}^2]$. Condition A.3 makes both strictly positive on $[c_L, c_U]$. □

Corollary A.1 (Existence and uniqueness of the calibrated scale)**.** *Let* $\rho(c)$ *denote either* $\tilde{\rho}(c)$ *or* $\bar{w}(c)$*. If* $\rho(c)$ *is continuous and strictly increasing on* $[c_L, c_U]$*, then for any target* $\rho^* \in (\rho(c_L), \rho(c_U))$ *there exists a unique* $c^* \in (c_L, c_U)$ *such that* $\rho(c^*) = \rho^*$*.*

Proof. Continuity follows from Conditions A.1–A.2 and continuity of the mapping $c \mapsto \mathcal{J}(\theta; c)$. Strict monotonicity is Proposition A.1. The result follows from the intermediate value theorem and injectivity of a strictly increasing function. □

A.4. **Jensen's Inequality and the Reliability Estimand Gap.** Section 2.2 notes that $\tilde{\rho}(c)$ is typically larger than $\bar{w}(c)$. We provide the formal argument and its implication for calibrated scales.

A.4.1. *Proof of* $\tilde{\rho}(c) \geq \bar{w}(c)$.

Proposition A.2 (Jensen inequality for reliability)**.** *Under Condition A.1,*

$$\tilde{\rho}(c) \geq \bar{w}(c) \qquad \text{for all } c \in [c_L, c_U]. \tag{A.16}$$

Equality holds iff $\mathcal{J}(\theta; c)$ *is* G*-a.s. constant.*

Proof. The function $x \mapsto 1/x$ is convex on $(0, \infty)$. By Jensen's inequality,

$$\mathbb{E}\left[\frac{1}{\mathcal{J}(\theta; c)}\right] \geq \frac{1}{\mathbb{E}[\mathcal{J}(\theta; c)]} = \frac{1}{\bar{\mathcal{J}}(c)}. \tag{A.17}$$

Using Equations (A.5) and (A.6),

$$\bar{w}(c) = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \mathbb{E}[1/\mathcal{J}]} \leq \frac{\sigma_\theta^2}{\sigma_\theta^2 + 1/\bar{\mathcal{J}}(c)} = \frac{\sigma_\theta^2 \bar{\mathcal{J}}(c)}{\sigma_\theta^2 \bar{\mathcal{J}}(c) + 1} = \tilde{\rho}(c).$$

Equality in Jensen holds iff $1/\mathcal{J}(\theta; c)$ is a.s. constant, equivalently $\mathcal{J}(\theta; c)$ is a.s. constant. □

A.4.2. *Implication for calibrated scales across estimands.* Let $c^*_{\tilde{\rho}}$ solve $\tilde{\rho}(c) = \rho^*$ and $c^*_{\bar{w}}$ solve $\bar{w}(c) = \rho^*$ (when solutions exist and are unique on $[c_L, c_U]$).

Corollary A.2 (Scale ordering across estimands)**.** *If* $\bar{w}$ *is strictly increasing on* $[c_L, c_U]$ *and both roots exist, then*

$$c^*_{\bar{w}} \geq c^*_{\tilde{\rho}}. \tag{A.18}$$

Proof. By Proposition A.2, $\bar{w}(c) \leq \tilde{\rho}(c)$ for all c. In particular, $\bar{w}(c^*_{\tilde{\rho}}) \leq \tilde{\rho}(c^*_{\tilde{\rho}}) = \rho^*$. Since $\bar{w}$ is strictly increasing, achieving $\bar{w}(c) = \rho^*$ requires $c \geq c^*_{\tilde{\rho}}$. □

A.4.3. *A second-order characterization of the Jensen gap.* Write $J = \mathcal{J}(\theta; c)$ and $\mu = \mathbb{E}[J]$. If J has finite variance and is sufficiently concentrated around μ, a second-order Taylor expansion gives

$$\mathbb{E}\left[\frac{1}{J}\right] = \frac{1}{\mu} + \frac{\mathrm{Var}(J)}{\mu^3} + R_3, \tag{A.19}$$

where R_3 is a third-order remainder controlled by $\mathbb{E}[|J - \mu|^3]$. Plugging Equation (A.19) into Equation (A.5) shows that the gap $\tilde{\rho}(c) - \bar{w}(c)$ is governed, to second order, by $\mathrm{Var}\{\mathcal{J}(\theta; c)\}$: the more unevenly information is distributed across the latent population, the larger the arithmetic–harmonic discrepancy and thus the larger the estimand gap.

A related local expansion for the calibrated scale gap follows from an implicit-function argument:

$$c^*_{\bar{w}} - c^*_{\tilde{\rho}} \approx \frac{\tilde{\rho}(c^*_{\tilde{\rho}}) - \bar{w}(c^*_{\tilde{\rho}})}{\bar{w}'(c^*_{\tilde{\rho}})}, \tag{A.20}$$

whenever $\bar{w}'(c^*_{\tilde{\rho}}) > 0$.

A.4.4. *A simple lower bound highlighting MSEM sensitivity.* Let $A_\varepsilon(c) = \{\theta : \mathcal{J}(\theta; c) \le \inf_\vartheta \mathcal{J}(\vartheta; c) + \varepsilon\}$. Then

$$\mathrm{MSEM}(c) = \mathbb{E}\left[\frac{1}{\mathcal{J}(\theta; c)}\right] \ge \frac{G\{A_\varepsilon(c)\}}{\inf_\vartheta \mathcal{J}(\vartheta; c) + \varepsilon}, \qquad \varepsilon > 0. \tag{A.21}$$

This inequality makes explicit why MSEM-based reliability is particularly sensitive to low-information regions: even a small amount of G-mass placed where $\mathcal{J}(\theta; c)$ is very small can substantially inflate MSEM.

A.5. **EQC: Consistency and Asymptotic Behavior of the Calibrated Root.** Section 3.2 defines EQC by fixing an empirical quadrature $\{\theta_m\}_{m=1}^M$ (and a fixed test form Ψ) and solving for a root of the empirical reliability curve.

A.5.1. *Empirical quadrature objects.* Let $\{\theta_m\}_{m=1}^M$ be i.i.d. draws from G. For a fixed Ψ, define

$$\hat{\bar{\mathcal{J}}}_M(c) = \frac{1}{M}\sum_{m=1}^M \mathcal{J}(\theta_m; c), \qquad \hat{\rho}_M(c) = \frac{\sigma_\theta^2\, \hat{\bar{\mathcal{J}}}_M(c)}{\sigma_\theta^2\, \hat{\bar{\mathcal{J}}}_M(c) + 1}, \tag{A.22}$$

which corresponds to the default EQC implementation targeting $\tilde{\rho}$. (If EQC is configured to target $\bar{w}$, replace $\hat{\bar{\mathcal{J}}}_M$ by $\widehat{\mathrm{MSEM}}_M(c) = \frac{1}{M}\sum_m 1/\mathcal{J}(\theta_m; c)$ and apply Equation (A.5).)

Let $\hat{c}_M$ denote the EQC root:

$$\hat{c}_M \in [c_L, c_U] \quad \text{s.t.} \quad \hat{\rho}_M(\hat{c}_M) = \rho^*. \tag{A.23}$$

A.5.2. *Uniform convergence and root consistency.*

Theorem A.1 (EQC uniform convergence and root consistency)**.** *Assume Conditions A.1–A.3 and that the function class $\{\mathcal{J}(\cdot; c) : c \in [c_L, c_U]\}$ satisfies a uniform law of large numbers under G. Then*

$$\sup_{c\in[c_L,c_U]} |\hat{\rho}_M(c) - \tilde{\rho}(c)| \xrightarrow{a.s.} 0,$$

and if c^* is the unique solution to $\tilde{\rho}(c) = \rho^*$ on $[c_L, c_U]$, then

$$\hat{c}_M \xrightarrow{a.s.} c^*.$$

Proof sketch. Uniform convergence of $\hat{\bar{\mathcal{J}}}_M(c)$ to $\bar{\mathcal{J}}(c)$ implies uniform convergence of $\hat{\rho}_M(c) = r(\hat{\bar{\mathcal{J}}}_M(c))$ to $\tilde{\rho}(c) = r(\bar{\mathcal{J}}(c))$ by continuity of r. Since $\tilde{\rho}$ is continuous and strictly increasing on $[c_L, c_U]$, the root mapping is continuous under uniform perturbations, yielding $\hat{c}_M \to c^*$. □

A.5.3. *Asymptotic normality (fixed test form).*

Theorem A.2 (Asymptotic normality of $\hat{c}_M$; fixed Ψ)**.** *Suppose, in addition to Theorem A.1, that $\tilde{\rho}$ is differentiable at c^* with $\tilde{\rho}'(c^*) > 0$, and $\mathrm{Var}\{\mathcal{J}(\theta; c^*)\} < \infty$. Then*

$$\sqrt{M}\,(\hat{c}_M - c^*) \xrightarrow{d} \mathcal{N}\left(0,\ \frac{\mathrm{Var}\{\mathcal{J}(\theta; c^*)\}}{\{\bar{\mathcal{J}}'(c^*)\}^2}\right). \tag{A.24}$$

Proof sketch. Since calibrating $\tilde{\rho}$ is equivalent to calibrating $\bar{\mathcal{J}}(c)$ to a fixed target value (because r is one-to-one), the result follows from a CLT for $\hat{\bar{\mathcal{J}}}_M(c^*)$ and a standard delta method for roots (one-dimensional M-estimation). □

A.6. **SAC: Robbins–Monro Convergence and Polyak–Ruppert Averaging.** Section 3.3 defines SAC as a Robbins–Monro procedure with projection and Polyak–Ruppert averaging. We state standard sufficient conditions for convergence in the present setting.

A.6.1. *Stochastic approximation form.* Let $\rho(c)$ denote the target reliability functional (typically $\bar{w}$ in SAC). Define

$$g(c) = \rho(c) - \rho^*, \qquad \text{so that} \qquad g(c^*) = 0. \tag{A.25}$$

At iteration n, SAC forms a noisy estimate $\hat{\rho}_n$ of $\rho(c_n)$ via fresh Monte Carlo sampling (Algorithm Box 2) and updates

$$c_{n+1} = \Pi_{[c_L,c_U]}\Big[c_n - a_n(\hat{\rho}_n - \rho^*)\Big] = \Pi_{[c_L,c_U]}\Big[c_n - a_n\{g(c_n) + \xi_{n+1}\}\Big], \tag{A.26}$$

where $\xi_{n+1} = \hat{\rho}_n - \rho(c_n)$ and Π denotes projection onto $[c_L, c_U]$. The step size sequence (Section 3.3) is

$$a_n = \frac{a}{(n + A)^\gamma}, \qquad a > 0,\ A \geq 0,\ \gamma \in (1/2, 1]. \tag{A.27}$$

A.6.2. *Almost sure convergence.*

Theorem A.3 (SAC convergence; standard Robbins–Monro conditions)**.** *Assume:*

(1) $a_n > 0$, $\sum_n a_n = \infty$, and $\sum_n a_n^2 < \infty$ (satisfied by Equation (A.27) with $\gamma \in (1/2, 1]$);

(2) $c^* \in (c_L, c_U)$ and projection as in Equation (A.26) is used;

(3) g *is continuous on* $[c_L, c_U]$ *and strictly increasing with* $g(c) < 0$ *for* $c < c^*$ *and* $g(c) > 0$ *for* $c > c^*$*;*

(4) $\{\xi_n\}$ *is a martingale-difference sequence w.r.t. the natural filtration* $\mathcal{F}_n$*, i.e.,*

$$\mathbb{E}[\xi_{n+1} \mid \mathcal{F}_n] = 0, \qquad \sup_n \mathbb{E}[\xi_{n+1}^2 \mid \mathcal{F}_n] < \infty \quad \text{a.s.} \tag{A.28}$$

Then

$$c_n \xrightarrow{a.s.} c^*. \tag{A.29}$$

Proof sketch. This is a direct application of classical stochastic approximation with projection. Condition (3) ensures a globally stable root on $[c_L, c_U]$ (ODE method), Condition (4) controls stochastic fluctuations, and Condition (1) guarantees diminishing adaptation noise with persistent excitation. Projection provides stability (bounded iterates). □

A.6.3. *Polyak–Ruppert averaging and warm-start intuition.* Define the post–burn-in Polyak–Ruppert average (Section 3.3):

$$\bar{c}_N = \frac{1}{N - B} \sum_{n=B+1}^{N} c_n. \tag{A.30}$$

Under standard differentiability and local moment conditions (e.g., g differentiable at c^* with $g'(c^*) > 0$), Polyak–Ruppert averaging yields the optimal $\sqrt{N}$-rate:

$$\sqrt{N}\,(\bar{c}_N - c^*) \xrightarrow{d} \mathcal{N}\left(0,\ \frac{\mathbb{V}}{g'(c^*)^2}\right), \tag{A.31}$$

where $\mathbb{V}$ is the asymptotic variance of the noise process at stationarity (i.e., the variability of the Monte Carlo reliability estimator near c^*).

Remark 2 (Why EQC warm start helps)**.** With diminishing step sizes, early SAC iterates can have disproportionate influence on the finite-N average $\bar{c}_N$. Initializing SAC at $c_0 \approx c^*$ (e.g., $c_0 = c^*_{\text{EQC}}$) reduces the transient regime, thereby shortening the burn-in required for stable averaging.

A.7. **Summary of What Appendix A Establishes.**

- Closed-form derivative identities for item/test information under global discrimination scaling (Equations (A.12) and (A.13)).
- Derivatives for both reliability functionals (Equations (A.14) and (A.15)) and the logic that reduces monotonicity to positivity of specific derivative expectations on a practical interval $[c_L, c_U]$ (Condition A.3).

- Existence and uniqueness of the calibrated scale c^* for targets within $(\rho(c_L), \rho(c_U))$ (Corollary A.1).
- Jensen inequality ordering $\tilde{\rho}(c) \geq \bar{w}(c)$ (Proposition A.2), implying $c^*_{\bar{w}} \geq c^*_{\tilde{\rho}}$ when calibrating to the same numerical target (Corollary A.2).
- Asymptotic justification for the calibration algorithms: EQC root consistency and a $\sqrt{M}$ limit under a fixed test form (Theorems A.1 and A.2), and SAC almost sure convergence with $\sqrt{N}$-rate under Polyak–Ruppert averaging (Theorem A.3 and Equation (A.31)).

Appendix B. Achievable Reliability Bounds

This appendix provides a rigorous discussion of achievable reliability bounds for reliability-targeted IRT simulation. It expands Section 2.4 of the main text by (i) formalizing feasibility on a practical calibration interval $[c_L, c_U]$, (ii) giving analytic upper bounds for the average-information reliability $\tilde{\rho}(c)$, (iii) characterizing intrinsic limitations of the MSEM-based reliability $\bar{w}(c)$ under information gaps, and (iv) outlining an empirical illustration roadmap (with figure specifications) to support interpretation and reproducibility.

Throughout, we use the notation of Appendix A. Fix an item configuration

$$\Psi = \{(\beta_i, \lambda_{i,0})\}_{i=1}^{I}, \qquad \lambda_{i,0} > 0,$$

and apply global discrimination scaling $\lambda_i(c) = c\lambda_{i,0}$ (Equation (2.5)). Let $\mathcal{J}(\theta; c)$ denote test information (Equation (2.2)), and let $\rho(c)$ denote either of the two population reliability functionals:

$$\tilde{\rho}(c) = \frac{\sigma_\theta^2 \bar{\mathcal{J}}(c)}{\sigma_\theta^2 \bar{\mathcal{J}}(c) + 1} \text{ (Equation (2.4))}, \qquad \bar{w}(c) = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \mathrm{MSEM}(c)} \text{ (Equation (2.3))},$$

with $\bar{\mathcal{J}}(c) = \mathbb{E}[\mathcal{J}(\theta; c)]$ and $\mathrm{MSEM}(c) = \mathbb{E}[1/\mathcal{J}(\theta; c)]$.

B.1. **Achievable Reliability Set, Endpoint Bounds, and Feasibility.** For a fixed (Ψ, G) and a fixed calibration interval $[c_L, c_U] \subset (0, \infty)$, define the achievable reliability set

$$\mathcal{R}([c_L, c_U]) \equiv \{\rho(c) : c \in [c_L, c_U]\}. \tag{B.1}$$

In complete generality (i.e., without any monotonicity assumption), define the global attainable extrema on $[c_L, c_U]$ as

$$\rho^\star_{\min} \equiv \inf_{c \in [c_L, c_U]} \rho(c), \qquad \rho^\star_{\max} \equiv \sup_{c \in [c_L, c_U]} \rho(c). \tag{B.2}$$

These are the appropriate mathematical notions of "achievable bounds" when $\rho(c)$ may be non-monotone on $[c_L, c_U]$.

In the main text (Section 2.4), we work in the practical calibration regime (Appendix A, Conditions A.1–A.3), where $\rho(c)$ is continuous and strictly increasing on $[c_L, c_U]$. In that regime, the attainable extrema simplify to endpoint values.

Proposition B.1 (Interval property under monotonicity)**.** *If $\rho(c)$ is continuous and strictly increasing on $[c_L, c_U]$, then*

$$\mathcal{R}([c_L, c_U]) = [\rho(c_L), \rho(c_U)], \qquad \rho^\star_{\min} = \rho(c_L), \quad \rho^\star_{\max} = \rho(c_U). \tag{B.3}$$

Proof. Continuity implies $\rho([c_L, c_U])$ is an interval. Strict monotonicity implies its minimum and maximum are attained at c_L and c_U, respectively. □

Feasibility of the inverse design problem. Under Proposition B.1, the inverse problem "find c^* such that $\rho(c^*) = \rho^*$" is feasible with a unique solution if and only if

$$\rho(c_L) < \rho^* < \rho(c_U). \tag{B.4}$$

If monotonicity fails on $[c_L, c_U]$ (which may occur for $\bar{w}$ under extreme scaling; see Section B.3), then Equation (B.4) is no longer a sufficient characterization of feasibility or uniqueness, and the inverse map can admit multiple solutions or none.

Remark 3 (IRTsimrel feasibility diagnostics)**.** The EQC implementation in IRTsimrel reports boundary reliabilities $\rho(c_L)$ and $\rho(c_U)$ to enable direct feasibility screening prior to interpreting the calibrated solution. In the package output these appear as `misc$rho_bounds["rho_L"]` and `misc$rho_bounds["rho_U"]`.

B.2. **Analytic Upper Bounds and Asymptotics for $\tilde{\rho}(c)$.** This section develops universal analytic bounds for $\tilde{\rho}(c)$ based on the logistic variance kernel. These bounds hold for any fixed (Ψ, G) and help interpret how test length and baseline discriminations constrain the upper tail of achievable $\tilde{\rho}$ values on a fixed $[c_L, c_U]$.

B.2.1. *A sharp bound for the logistic variance kernel.* Let $s(x) = (1 + e^{-x})^{-1}$ and $h(x) = s(x)\{1 - s(x)\}$. Under global scaling, item information can be written (Appendix A, Equation (A.11)) as

$$\mathcal{J}_i(\theta; c) = c^2 \lambda_{i,0}^2 h\left(c\lambda_{i,0}(\theta - \beta_i)\right). \tag{B.5}$$

Lemma B.1 (Kernel bound)**.** *For all $x \in \mathbb{R}$,*

$$0 < h(x) \le \frac{1}{4}, \qquad \textit{with equality iff } x = 0. \tag{B.6}$$

Proof. $h(x) = p(1-p)$ with $p = s(x) \in (0,1)$, maximized at $p = 1/2$ (i.e., $x = 0$) with maximum $1/4$. □

B.2.2. *Bounds for test information and $\tilde{\rho}(c)$.* Summing Equation (B.5) and applying Lemma B.1 yields a pointwise test-information bound.

Proposition B.2 (Pointwise and average information upper bounds)**.** *For any $c > 0$ and any $\theta \in \mathbb{R}$,*

$$\mathcal{J}(\theta; c) = \sum_{i=1}^{I} c^2 \lambda_{i,0}^2 h\left(c\lambda_{i,0}(\theta - \beta_i)\right) \leq \frac{c^2}{4} \sum_{i=1}^{I} \lambda_{i,0}^2. \tag{B.7}$$

Consequently,

$$\bar{\mathcal{J}}(c) = \mathbb{E}\left[\mathcal{J}(\theta; c)\right] \leq \frac{c^2}{4} \sum_{i=1}^{I} \lambda_{i,0}^2. \tag{B.8}$$

Proof. Apply Equation (B.6) termwise in Equation (B.5) and sum; then take expectations. □

Plugging Equation (B.8) into the definition of $\tilde{\rho}(c)$ (Equation (2.4)) gives an explicit ceiling.

Corollary B.1 (Closed-form ceiling for $\tilde{\rho}(c)$)**.** *Let $S_2 \equiv \sum_{i=1}^{I} \lambda_{i,0}^2$. For any $c > 0$,*

$$\tilde{\rho}(c) = \frac{\sigma_\theta^2 \bar{\mathcal{J}}(c)}{\sigma_\theta^2 \bar{\mathcal{J}}(c) + 1} \leq \frac{\sigma_\theta^2 (c^2 S_2/4)}{\sigma_\theta^2 (c^2 S_2/4) + 1}. \tag{B.9}$$

Special case (Rasch, standardized G). If $\lambda_{i,0} \equiv 1$ and $\sigma_\theta^2 = 1$, then $S_2 = I$ and

$$\tilde{\rho}(c) \leq \frac{c^2 I/4}{c^2 I/4 + 1}. \tag{B.10}$$

Evaluating Equation (B.10) at $c = c_U$ yields a simple analytic upper bound on the achievable $\tilde{\rho}$ range induced by the calibration cap $c \leq c_U$.

B.2.3. *Small-c behavior (quadratic onset).* As $c \downarrow 0$, $\pi_i(\theta; c) \to 1/2$ and $h(c\lambda_{i,0}(\theta - \beta_i)) \to 1/4$ for each fixed θ. For fixed Ψ, this implies

$$\mathcal{J}(\theta; c) = \frac{c^2}{4} \sum_{i=1}^{I} \lambda_{i,0}^2 + o(c^2), \qquad c \downarrow 0, \tag{B.11}$$

and hence $\tilde{\rho}(c) = O(c^2)$ near the origin. Practically, once c_L is already small, further reducing c_L expands $\tilde{\rho}(c_L)$ only at a quadratic rate.

B.2.4. *Large-c behavior for $\tilde{\rho}(c)$ (linear growth of $\bar{\mathcal{J}}(c)$).* A key point for interpretation is that $\tilde{\rho}(c)$ depends on the arithmetic mean $\bar{\mathcal{J}}(c)$, and therefore can approach 1 even when information becomes highly uneven across θ.

Assume G has a continuous density g. For a fixed item i,

$$\mathbb{E}\big[\mathcal{J}_i(\theta; c)\big] = \int c^2 \lambda_{i,0}^2 h\left(c\lambda_{i,0}(\theta - \beta_i)\right) g(\theta) d\theta. \tag{B.12}$$

Let $u = c\lambda_{i,0}(\theta - \beta_i)$, so $d\theta = du/(c\lambda_{i,0})$. Then

$$\mathbb{E}\big[\mathcal{J}_i(\theta; c)\big] = c\lambda_{i,0} \int h(u) g\left(\beta_i + \frac{u}{c\lambda_{i,0}}\right) du. \tag{B.13}$$

Because $h(u) = s'(u)$ is the logistic density, $\int_{-\infty}^{\infty} h(u)du = 1$. Under dominated convergence (using integrability of h and continuity of g),

$$\mathbb{E}\big[\mathcal{J}_i(\theta; c)\big] \sim c\lambda_{i,0} g(\beta_i), \qquad c \to \infty. \tag{B.14}$$

Summing over items yields

$$\bar{\mathcal{J}}(c) \sim c \sum_{i=1}^{I} \lambda_{i,0} g(\beta_i), \qquad c \to \infty, \tag{B.15}$$

and therefore $\tilde{\rho}(c) \to 1$ as $c \to \infty$ whenever $\sum_i \lambda_{i,0} g(\beta_i) > 0$. This asymptotic clarifies why, for $\tilde{\rho}$, the effective upper bound in practice is usually imposed by the chosen c_U, rather than by an intrinsic ceiling strictly below 1.

B.3. **Intrinsic Limitations of $\bar{w}(c)$ Under Information Gaps.** Unlike $\tilde{\rho}(c)$, the MSEM-based reliability depends on the harmonic mean of information through $\mathbb{E}[1/\mathcal{J}(\theta; c)]$:

$$\bar{w}(c) = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \mathbb{E}[1/\mathcal{J}(\theta; c)]}. \tag{B.16}$$

This structure makes $\bar{w}(c)$ highly sensitive to low-information regions. Appendix A formalizes the Jensen-inequality ordering $\tilde{\rho}(c) \geq \bar{w}(c)$ (Proposition A.2), but here we emphasize a distinct phenomenon: increasing discrimination can worsen $\bar{w}(c)$ if it creates deep information gaps between increasingly "spiky" item-information peaks.

B.3.1. *Collapse under a genuine difficulty gap.* The following result formalizes an extreme but instructive mechanism: if there is a region of non-negligible latent mass with no nearby item difficulties, then $\bar{w}(c)$ cannot be driven upward by arbitrarily large c; it can in fact collapse.

Proposition B.3 (MSEM explosion and $\bar{w}(c)$ collapse under a difficulty gap)**.** *Assume:*

(1) G has a density g, and there exists an interval $[a, b]$ such that $\int_a^b g(\theta)d\theta > 0$.

(2) There exists $\delta > 0$ such that for all $\theta \in [a, b]$ and all i, $|\theta - \beta_i| \geq \delta$ (no item difficulty lies within distance δ of $[a, b]$).

(3) Let $\lambda_{\min} \equiv \min_i \lambda_{i,0} > 0$, and define $S_2 = \sum_i \lambda_{i,0}^2$.

Then, as $c \to \infty$,

$$\mathrm{MSEM}(c) = \mathbb{E}\left[\frac{1}{\mathcal{J}(\theta; c)}\right] \to \infty \quad \textit{and hence} \quad \bar{w}(c) \to 0. \tag{B.17}$$

Proof. For $x \geq 0$, $h(x) = e^{-x}/(1 + e^{-x})^2 \leq e^{-x}$; by symmetry, $h(x) \leq e^{-|x|}$ for all $x \in \mathbb{R}$. For any $\theta \in [a, b]$, condition (2) gives $|\theta - \beta_i| \geq \delta$ for all i, so

$$\mathcal{J}(\theta; c) = \sum_{i=1}^{I} c^2 \lambda_{i,0}^2 h\left(c\lambda_{i,0}(\theta - \beta_i)\right) \leq c^2 \sum_{i=1}^{I} \lambda_{i,0}^2 \exp\{-c\lambda_{i,0}|\theta - \beta_i|\} \leq c^2 S_2 \exp\{-c\lambda_{\min}\delta\}. \tag{B.18}$$

Therefore, for $\theta \in [a, b]$,

$$\frac{1}{\mathcal{J}(\theta; c)} \geq \frac{\exp\{c\lambda_{\min}\delta\}}{c^2 S_2}. \tag{B.19}$$

Integrating over $[a, b]$ yields

$$\mathrm{MSEM}(c) = \int \frac{1}{\mathcal{J}(\theta; c)} g(\theta)d\theta \geq \left(\int_a^b g(\theta)d\theta\right) \frac{\exp\{c\lambda_{\min}\delta\}}{c^2 S_2} \to \infty, \tag{B.20}$$

which implies $\bar{w}(c) \to 0$. □

Remark 4 (Connection to "practical calibration intervals")**.** Proposition B.3 is a large-c pathology: it demonstrates that $\bar{w}(c)$ need not be globally increasing on $(0, \infty)$, even though it is strictly increasing on a practical $[c_L, c_U]$ under Appendix A, Condition A.3. In applied calibration, one should therefore interpret c_U not merely as a numerical convenience, but as a modeling choice that restricts attention to the regime where the inverse design problem remains well-posed for $\bar{w}$.

B.4. **Conservative Ceilings and Feasibility Screening for Design.** This section summarizes how the preceding results translate into concrete feasibility checks and target-grid construction in simulation design.

B.4.1. *A computable analytic ceiling for $\tilde{\rho}$ on $[c_L, c_U]$.* For $\tilde{\rho}$, Corollary B.1 provides a direct analytic upper bound valid for any (Ψ, G). In particular, for any calibration cap $c \leq c_U$,

$$\sup_{c \in [c_L, c_U]} \tilde{\rho}(c) = \tilde{\rho}(c_U) \leq \frac{\sigma_\theta^2(c_U^2 S_2/4)}{\sigma_\theta^2(c_U^2 S_2/4) + 1}, \qquad S_2 = \sum_{i=1}^{I} \lambda_{i,0}^2, \tag{B.21}$$

where equality in the first step uses monotonicity on $[c_L, c_U]$ (Appendix A).

This bound is useful as a quick impossibility check: if a proposed target ρ^* exceeds the right-hand side of Equation (B.21), then it is infeasible for $\tilde{\rho}$ on the chosen interval regardless of the specific difficulty locations.

B.4.2. *Back-of-envelope "reference-scale" ceilings for target-grid sanity checks.* In practice, researchers often wish to avoid targets that, while feasible in principle, would require extreme scaling and thus risk entering the non-monotone regime for $\bar{w}$ (Section B.3) or producing numerically ill-conditioned calibration near the boundary (Section 2.4). A simple heuristic—used only for sanity checking target grids—is to combine the Rasch maximum item information (0.25) (attained at $\theta = \beta_i$ when $c = 1$) with $\sigma_\theta^2 = 1$, yielding

$$\tilde{\rho}^{\text{ref}}_{\max}(I) \equiv \frac{I/4}{I/4 + 1}. \tag{B.22}$$

This is not a sharp achievable bound for a given (Ψ, G), nor does it incorporate scaling $c \neq 1$. Its role is to flag target grids that are likely to require unusually aggressive scaling for short tests.

TABLE B.1. Reference-scale ceiling for $\tilde{\rho}$ under Rasch and standardized G

Test length (I)	$I/4$	$\tilde{\rho}^{\text{ref}}_{\max}(I)$
15	3.75	0.7895
30	7.50	0.8824
60	15.00	0.9375

Note. Values computed from Equation (B.22). This "ceiling" assumes information is near its per-item maximum across the latent population at $c = 1$, and is therefore best interpreted as a conservative target-selection heuristic, not as the achievable $\tilde{\rho}(c_U)$ implied by a chosen calibration interval.

B.4.3. *Recommended feasibility-screening workflow.* For a proposed design (Ψ, G, ρ^*) and a chosen metric ($\tilde{\rho}$ or $\bar{w}$), we recommend:

(1) Compute boundary reliabilities $\rho(c_L)$ and $\rho(c_U)$ under the chosen metric.
(2) Check feasibility under the practical-monotone assumption: require $\rho(c_L) < \rho^* < \rho(c_U)$.
(3) Avoid targets too close to either bound, because the inverse map is ill-conditioned near the boundary: small Monte Carlo perturbations in $\hat{\rho}(c)$ can induce large perturbations in $\hat{c}$.

(4) For $\bar{w}$, add a monotonicity diagnostic if c_U is large or the item grid is sparse: evaluate $\bar{w}(c)$ on a coarse grid of c values to confirm it is increasing on $[c_L, c_U]$. If non-monotonicity is detected, reduce c_U or revise the difficulty coverage.

(5) If infeasible, adjust one or more of: (i) test length (I), (ii) item pool quality/coverage (difficulty support), (iii) c_U (to enlarge the feasible interval only if monotonicity is preserved), or (iv) the reliability metric (noting $\tilde{\rho} \geq \bar{w}$; Appendix A).

Appendix C. Latent Distribution Specifications

This appendix documents the implementation of latent trait distributions G used in the validation study (Section 4). All distributions are generated via the `sim_latentG()` function in the `IRTsimrel` package.

C.1. **Pre-Standardization Principle.** A key design feature is *pre-standardization*: every built-in distribution shape is mathematically constructed to have mean 0 and variance 1 before any location-scale transformation is applied. The generated abilities follow:

$$\theta_p = \mu + \sigma \cdot z_p, \qquad z_p \sim G_0, \quad \mathbb{E}[z] = 0, \quad \mathrm{Var}(z) = 1. \tag{C.1}$$

This design ensures that changing the distributional shape does not inadvertently alter the scale, enabling clean comparisons across shapes while holding variance constant. When $\mu = 0$ and $\sigma = 1$ (the default), the generated θ values have exactly the target mean and variance regardless of the underlying shape.

C.2. **Validation Study Distributions.** The validation study (Section 4) employed four latent distribution shapes representing qualitatively distinct departures from normality. Table C.1 summarizes their mathematical construction and higher-order moments.

Table C.1. Validation Study Distribution Parameters

Shape	Construction	Formula	Param.	Skew.	Ex. Kurt.
Normal	Standard normal	$z \sim N(0,1)$	—	0.00	0.00
Bimodal	Symmetric mixture	$z = s\delta + \varepsilon$	$\delta = 0.8$	0.00	-1.09
Pos. Skewed	Std. Gamma	$z = (\Gamma_k - k)/\sqrt{k}$	$k = 4$	1.00	1.50
Heavy-Tailed	Std. Student-t	$z = t_\nu/\sqrt{\nu/(\nu-2)}$	$\nu = 5$	0.00	6.00

Note. All distributions are pre-standardized to have $\mathbb{E}[z] = 0$ and $\mathrm{Var}(z) = 1$. Skewness refers to the third standardized moment; excess kurtosis refers to the fourth standardized moment minus 3. For the bimodal distribution, $s \sim$ Rademacher(± 1) and $\varepsilon \sim N(0, 1-\delta^2)$.

C.2.1. *Mathematical details for each shape.* **Normal.** The baseline case serves as a benchmark: $z \sim N(0, 1)$. This is the conventional assumption in IRT and represents the null hypothesis of no distributional misspecification.

Bimodal. A symmetric two-component Gaussian mixture represents populations with two distinct subgroups. The construction

$$z = s \cdot \delta + \varepsilon, \qquad s \sim \text{Rademacher}(\pm 1), \quad \varepsilon \sim N(0, 1 - \delta^2), \tag{C.2}$$

places mixture components at $\pm\delta$ with common within-component variance $1 - \delta^2$. This ensures $\text{Var}(z) = \delta^2 + (1 - \delta^2) = 1$ by construction. With $\delta = 0.8$, the modes are clearly separated while maintaining unit variance. The excess kurtosis is $-2\delta^4/(1 - \delta^2 + \delta^4) \approx -1.09$, reflecting the platykurtic nature of bimodal distributions.

Positively Skewed. The standardized Gamma distribution

$$z = \frac{\Gamma(k, 1) - k}{\sqrt{k}} \tag{C.3}$$

has $\mathbb{E}[z] = 0$ and $\text{Var}(z) = 1$ for any shape parameter $k > 0$. With $k = 4$, the distribution has skewness $2/\sqrt{k} = 1$ and excess kurtosis $6/k = 1.5$. This represents positively selective samples (e.g., high-ability populations).

Heavy-Tailed. The standardized Student-t distribution

$$z = \frac{t_\nu}{\sqrt{\nu/(\nu - 2)}} \tag{C.4}$$

has $\text{Var}(z) = 1$ for $\nu > 2$. With $\nu = 5$ degrees of freedom, the distribution has excess kurtosis $6/(\nu - 4) = 6$, representing substantially heavier tails than the normal. This shape is useful for examining robustness to outliers and extreme values.

Figure C.1 displays the density functions for all four validation study shapes, with the standard normal reference shown as a dashed curve.

C.3. **Implementation in IRTsimrel.** The `sim_latentG()` function generates latent abilities with the following interface:

```
sim_latentG(
  n,                  # Number of persons
  shape = "normal",   # Distribution shape
  shape_params = list(), # Shape-specific parameters
  mu = 0,             # Location parameter
  sigma = 1,          # Scale parameter
  seed = NULL         # Random seed for reproducibility
```

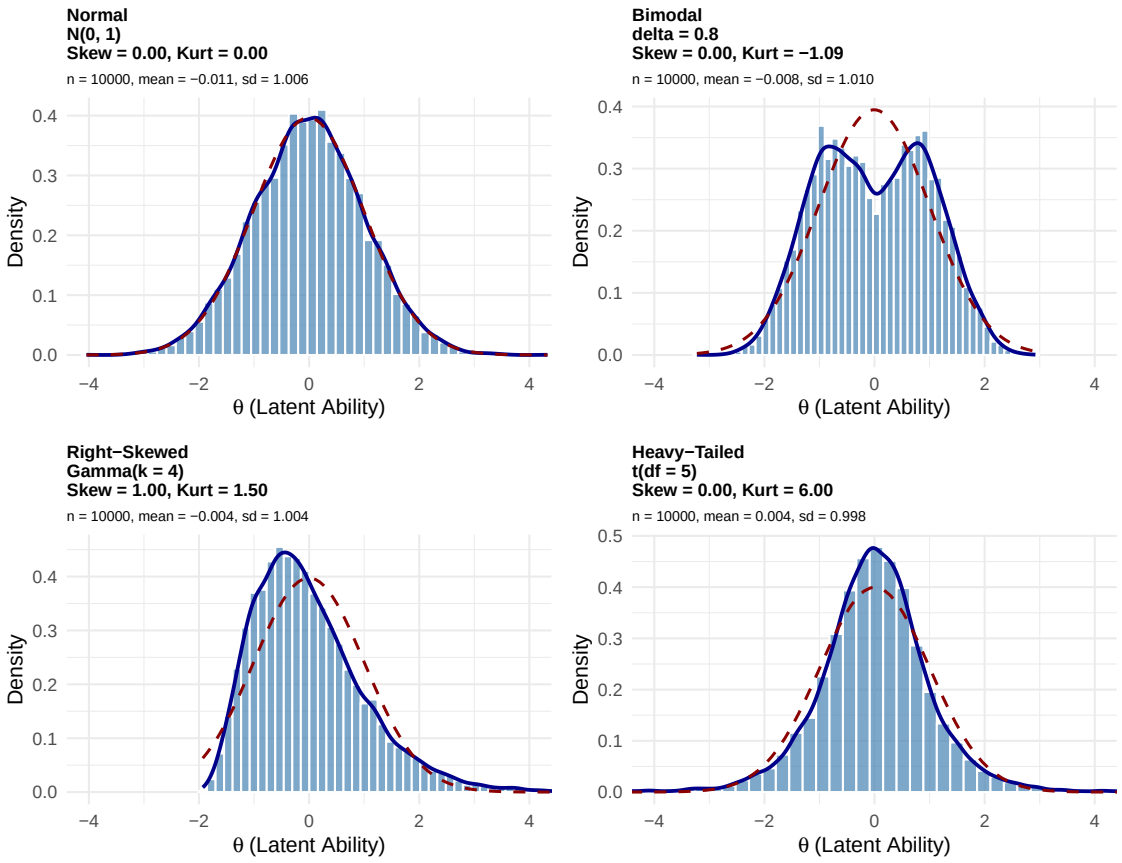

FIGURE C.1. Latent Distribution Shapes Used in Validation Study

Note. All distributions are pre-standardized to have mean 0 and variance 1. Solid blue curves show kernel density estimates from $n = 10{,}000$ draws; dashed red curves show the $N(0, 1)$ reference. Each panel reports the theoretical skewness (Skew) and excess kurtosis (Kurt), along with empirical sample moments (μ, σ) confirming the pre-standardization property.

)

For the validation study conditions, the function calls were:

```
# Normal
sim_latentG(n = N, shape = "normal", seed = seed)


# Bimodal
sim_latentG(n = N, shape = "bimodal",
            shape_params = list(delta = 0.8), seed = seed)
```

```
# Positively Skewed
sim_latentG(n = N, shape = "skew_pos",
            shape_params = list(k = 4), seed = seed)

# Heavy-Tailed
sim_latentG(n = N, shape = "heavy_tail",
            shape_params = list(df = 5), seed = seed)
```

The function returns a `latent_G` object containing the generated θ values, the pre-standardized z values, and sample moment diagnostics.

Appendix D. Item Parameter Generation Details

This appendix documents the generation of item parameters (difficulties β and discriminations λ) for the validation study. All item generation is performed via the `sim_item_params()` function in the `IRTsimrel` package.

D.1. **Item-Generation Configurations.** The validation study employed a 2×2 factorial design crossing IRT model (Rasch vs. 2PL) with difficulty source (parametric vs. empirical). Table D.1 summarizes the four configurations.

TABLE D.1. Item-Generation Configurations

Model	Source	Difficulty	Discrimination	Method
Rasch	Parametric	$\beta \sim N(0,1)$	$\lambda \equiv 1$ (fixed)	—
Rasch	IRW	IRW pool (empirical)	$\lambda \equiv 1$ (fixed)	—
2PL	Parametric	$\beta \sim N(0,1)$	$\log(\lambda) \sim N(0, 0.3^2)$, $\rho = -0.3$	Gaussian Copula
2PL	IRW	IRW pool (empirical)	$\log(\lambda) \sim N(0, 0.3^2)$, $\rho = -0.3$	Gaussian Copula

Note. IRW = Item Response Warehouse (Domingue et al., 2025). The correlation $\rho = -0.3$ reflects the empirically observed negative relationship between difficulty and discrimination (Sweeney et al., 2022).

D.2. **Difficulty Sources. Parametric source.** Difficulties are drawn from a standard normal distribution:

$$\beta_i \sim N(0,1), \qquad i = 1, \ldots, I. \tag{D.1}$$

This represents the conventional assumption in IRT simulation studies.

IRW source. Difficulties are sampled from the Item Response Warehouse (Domingue et al., 2025; Zhang et al., 2025), an empirical repository of calibrated item parameters

from operational assessments. The IRW pool exhibits a characteristically bimodal distribution with modes near $\beta \approx -2$ and $\beta \approx 1$, reflecting the structure of real item banks. Figure D.1 compares the parametric and IRW difficulty distributions.

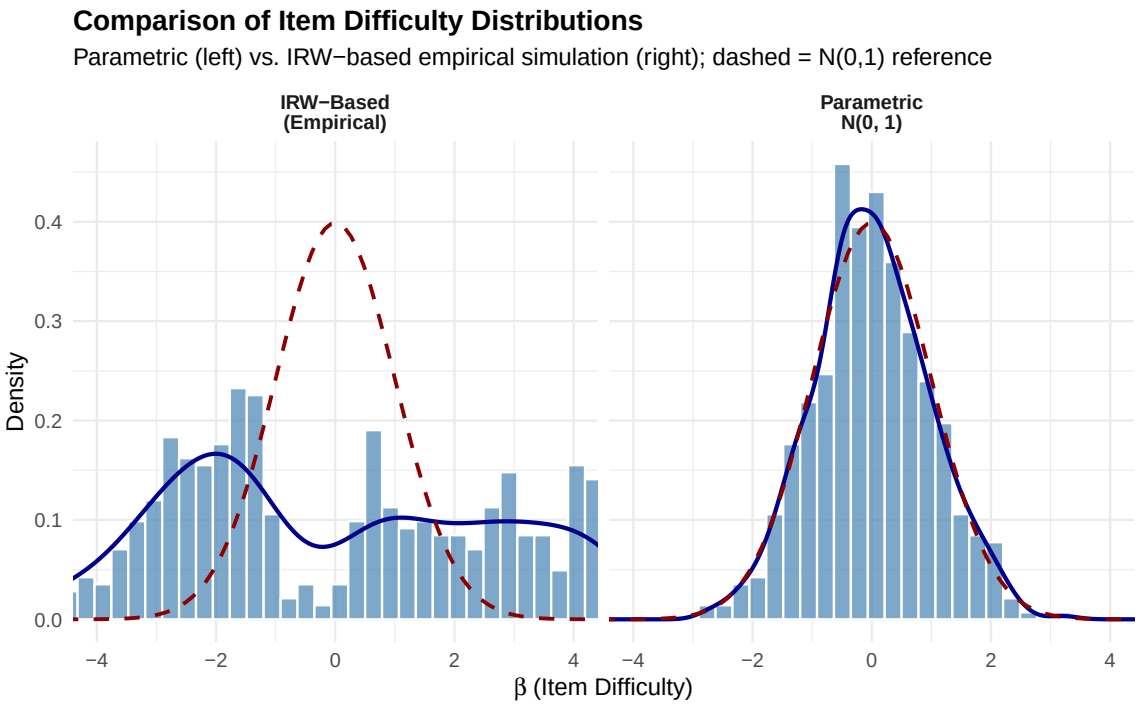

FIGURE D.1. Comparison of Item Difficulty Distributions

Note. Left panel: IRW-based empirical difficulties showing the characteristic bimodal structure of real assessment item pools. Right panel: Parametric $N(0, 1)$ difficulties. Solid blue curves show kernel density estimates; dashed red curves show the $N(0, 1)$ reference. The IRW distribution has substantially greater variance (SD ≈ 1.6) and pronounced non-normality compared to the parametric distribution.

D.3. **Discrimination Generation: The Gaussian Copula Method.** For the 2PL model, discriminations must be generated with a specified correlation to difficulties. A critical finding from psychometric research is that item difficulty and discrimination are negatively correlated in real assessments, with typical values around $\rho \approx -0.3$ (Sweeney et al., 2022). Ignoring this dependence produces unrealistic simulation data.

The `IRTsimrel` package implements a *Gaussian copula* method that achieves target correlations while exactly preserving the marginal distributions of both parameters. This is crucial when difficulties come from the non-normal IRW distribution.

Algorithm. Given a vector of difficulties $\{\beta_i\}_{i=1}^{I}$ from any source (parametric or IRW), the copula method generates correlated log-normal discriminations through the steps shown in Table D.2.

TABLE D.2. Gaussian Copula Algorithm Steps

Step	Operation	Formula	Purpose
1	Transform β to uniform	$u = \text{rank}(\beta)/(n+1)$	Nonparametric CDF
2	Transform uniform to normal	$z_\beta = \Phi^{-1}(u)$	Std. normal quantile
3	Generate correlated normal	$z_\lambda = \rho z_\beta + \sqrt{1-\rho^2} z_{\text{indep}}$	Impose correlation
4	Transform normal to uniform	$v = \Phi(z_\lambda)$	Std. normal CDF
5	Transform uniform to log-normal	$\log(\lambda) = \mu + \sigma\Phi^{-1}(v)$	Target marginal

Note. $\Phi(\cdot)$ denotes the standard normal CDF and $\Phi^{-1}(\cdot)$ its inverse. Parameters $\mu = 0$ and $\sigma = 0.3$ yield the target log-normal marginal for discriminations.

The key insight is that Step 1 uses the *empirical* (rank-based) CDF rather than assuming any parametric form for the difficulty distribution. This nonparametric transformation ensures that the original difficulty marginal—whether normal, bimodal, or any other shape—is exactly preserved in the output.

Figure D.2 illustrates the copula algorithm step-by-step, showing how the IRW difficulty distribution is transformed through each stage while achieving the target correlation.

D.4. **Comparison of Generation Methods.** The `IRTsimrel` package provides three methods for generating correlated item parameters:

(1) **Copula method** (recommended): Preserves exact marginals, achieves target Spearman correlation
(2) **Conditional method**: Uses conditional normal regression; assumes linear relationships
(3) **Independent method**: Generates discriminations independently (no correlation)

Figure D.3 compares these methods with a target Spearman $\rho = -0.3$.

D.5. **Joint Distribution of Item Parameters.** Figure D.4 displays the complete joint distribution of item parameters for the 2PL + IRW configuration, showing the marginal distributions and their bivariate relationship.

D.6. **Summary Statistics.** Table D.3 reports summary statistics for the four item-generation configurations used in the validation study.

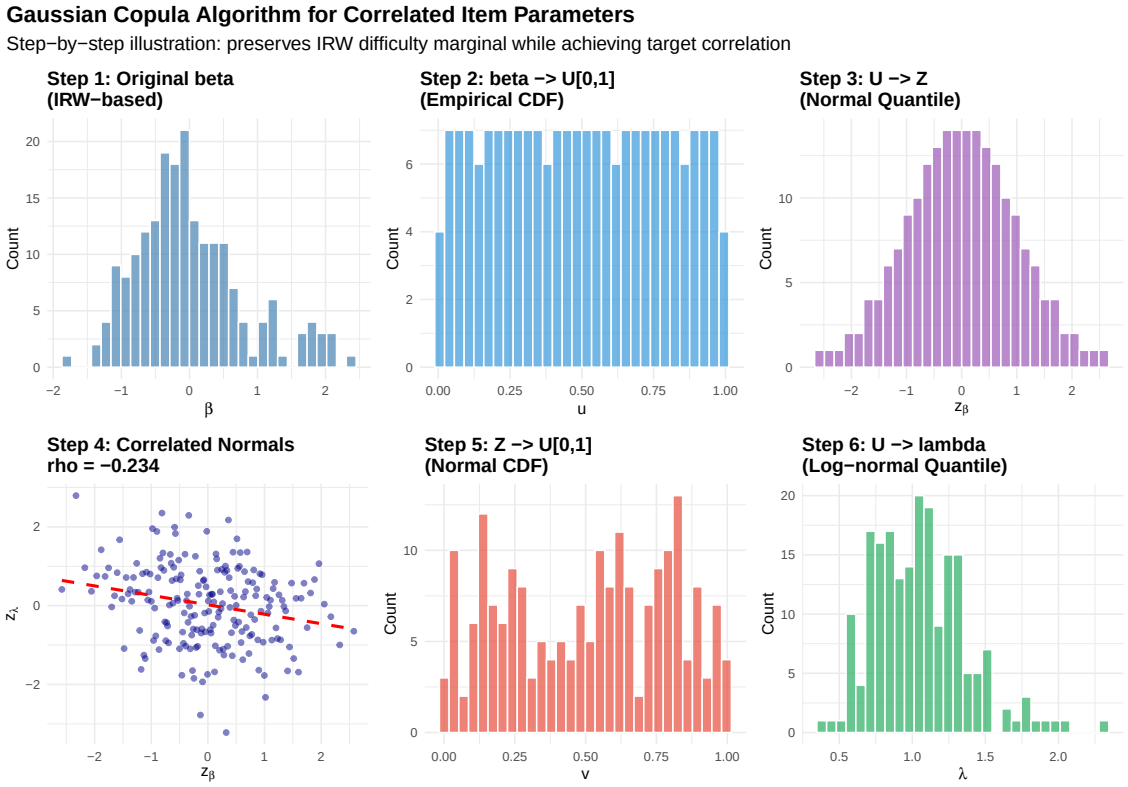

FIGURE D.2. Gaussian Copula Algorithm for Correlated Item Parameters

Note. Step-by-step illustration of the Gaussian copula method using $n = 200$ items from the IRW pool. Step 1: Original IRW difficulties. Steps 2–3: Transformation to uniform then normal space. Step 4: Generation of correlated normal pairs (achieved $\rho = -0.234$). Steps 5–6: Back-transformation through uniform to log-normal discriminations. The method preserves the exact IRW marginal while achieving the target Spearman correlation.

TABLE D.3. Item Parameter Summary Statistics

Configuration	$\bar{\beta}$	SD(β)	Range(β)	$\bar{\lambda}$	SD(λ)	r_S
Rasch + Parametric	0.00	0.97	$[-3.01, 3.60]$	1.00	0.00	—
Rasch + IRW	0.00	1.63	$[-3.11, 4.00]$	1.00	0.00	—
2PL + Parametric	0.00	1.00	$[-3.11, 3.09]$	1.05	0.34	−0.28
2PL + IRW	0.00	1.85	$[-3.68, 3.32]$	1.04	0.31	−0.29

Note. Statistics computed from $n = 1{,}000$ items per configuration. r_S = Spearman correlation between β and $\log(\lambda)$. IRW difficulties exhibit approximately 1.7× greater variability than parametric difficulties. Achieved correlations are within sampling error of the target $\rho = -0.3$.

D.7. **Implementation in IRTsimrel.** The `sim_item_params()` function generates item parameters with the following interface:

```
sim_item_params(
```

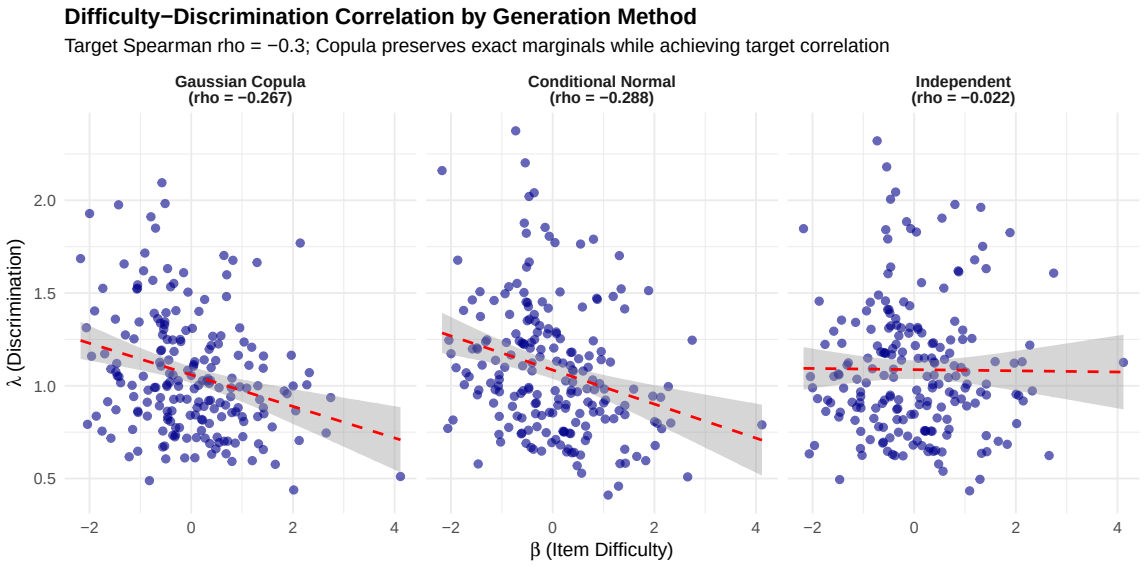

FIGURE D.3. Difficulty–Discrimination Correlation by Generation Method

Note. Comparison of three methods for generating correlated item parameters. Target Spearman $\rho = -0.3$. Left: Gaussian copula (achieved $\rho = -0.267$). Center: Conditional normal (achieved $\rho = -0.288$). Right: Independent generation (achieved $\rho = -0.022$). The copula method preserves the exact IRW difficulty marginal while achieving the target rank correlation.

```
  n_items,                    # Number of items
  model = "rasch",            # "rasch" or "2pl"
  source = "irw",             # "parametric", "irw", "hierarchical", or "custom"
  method = "copula",          # "copula", "conditional", or "independent"
  difficulty_params = list(), # Parameters for difficulty distribution
  discrimination_params = list(
    mu_log = 0,               # Mean of log(lambda)
    sigma_log = 0.3,          # SD of log(lambda)
    rho = -0.3                # Target correlation
  ),
  scale = 1,                  # Global discrimination scale factor
  seed = NULL                 # Random seed for reproducibility
)
```

For the validation study, item parameters were generated as:

```
# Rasch + Parametric
sim_item_params(n_items = I, model = "rasch", source = "parametric",
                difficulty_params = list(mu = 0, sigma = 1), seed = seed)
```

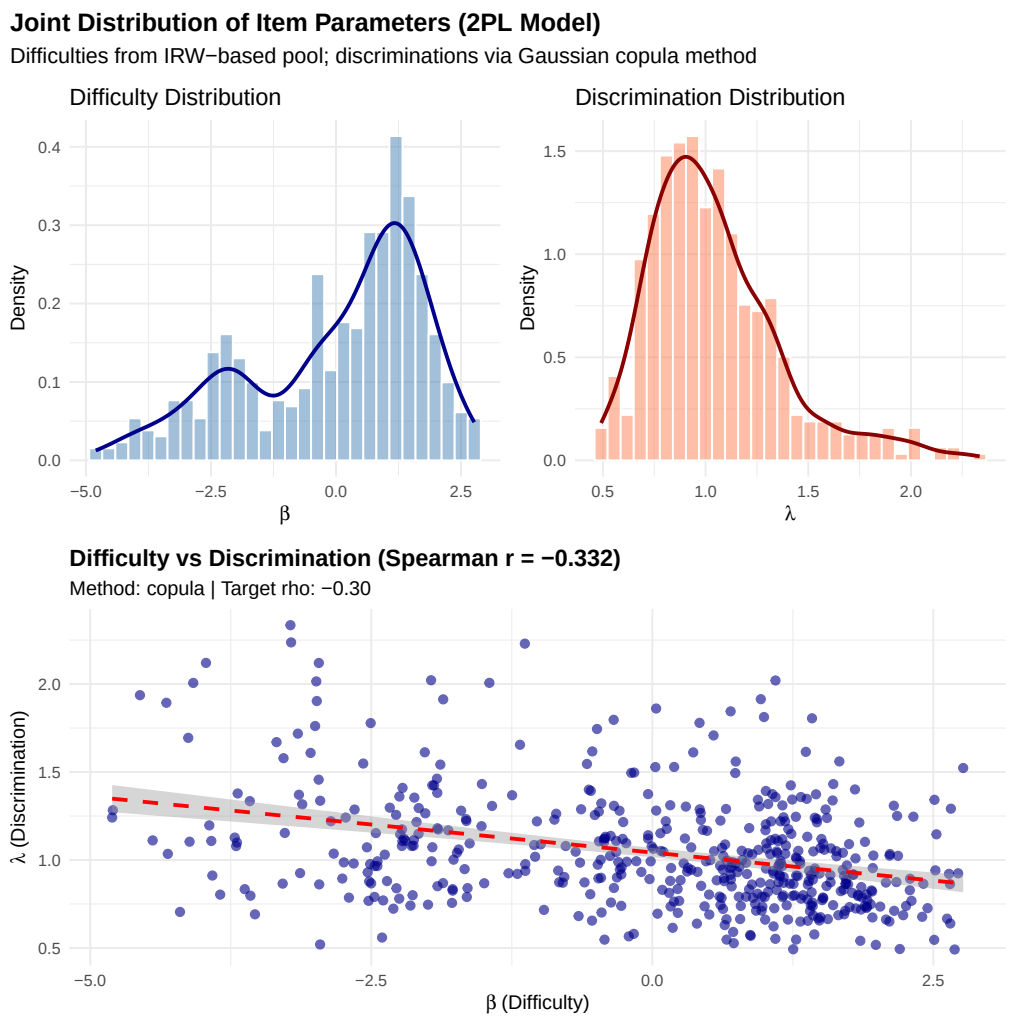

FIGURE D.4. Joint Distribution of Item Parameters (2PL Model)

Note. Joint distribution of item parameters generated via the Gaussian copula method with $n = 500$ items. Top left: Difficulty distribution from IRW pool showing characteristic bimodality. Top right: Log-normal discrimination distribution. Bottom: Scatterplot showing the negative correlation between difficulty and discrimination (Spearman $r = -0.332$; target $\rho = -0.30$). The copula method successfully imposes the target correlation while preserving both marginal distributions exactly.

```
# Rasch + IRW
sim_item_params(n_items = I, model = "rasch", source = "irw", seed = seed)

# 2PL + Parametric
sim_item_params(n_items = I, model = "2pl", source = "parametric",
```

```
                method = "copula",
                difficulty_params = list(mu = 0, sigma = 1),
                discrimination_params = list(mu_log = 0, sigma_log = 0.3,
                                             rho = -0.3),
                seed = seed)

# 2PL + IRW
sim_item_params(n_items = I, model = "2pl", source = "irw",
                method = "copula",
                discrimination_params = list(mu_log = 0, sigma_log = 0.3,
                                             rho = -0.3),
                seed = seed)
```

D.8. **Integration with Reliability Calibration.** In the reliability-targeted simulation framework, item parameters are generated with a baseline scale ($c = 1$) and then rescaled during calibration. The `eqc_calibrate()` function automatically calls `sim_item_params()` internally and applies the calibrated scaling factor c^*:

$$\lambda_i^* = c^* \cdot \lambda_{i,0}, \tag{D.2}$$

where $\lambda_{i,0}$ denotes the baseline discrimination. This separation of structure (the baseline parameters) from scale (the calibrated factor c^*) is central to the reliability-targeted simulation approach described in the main text (Section 2).

Appendix E. Extended Validation Results

This appendix provides extended validation results that complement the main text (Section 4). While Tables 2 and 3 and Figures 1 to 5 summarize aggregate calibration accuracy, the analyses presented here stratify results by design factors, quantify finite-sample replication variability, and provide additional empirical support for the theoretical properties established in Appendices A and B.

Throughout, we use the notation of the main text: $\tilde{\rho}$ denotes average-information reliability (Equation (2.4)), $\bar{w}$ denotes MSEM-based reliability (Equation (2.3)), and c^* denotes the calibrated discrimination scale. The validation study comprised 960 conditions crossing latent distribution shape (4 levels), IRT model (2 levels), item source (2 levels), test length $I \in \{15, 30, 60\}$, sample size $N \in \{100, 200, 500, 1000, 2000\}$, and target reliability ρ^* (4 levels per test length). For each condition, $K = 2{,}000$ replicated response datasets were generated to assess finite-sample variability.

E.1. **Calibration Accuracy by Test Length and Latent Distribution.** Figure 2 in the main text aggregated calibration deviations across latent distribution shapes. Figure E.1 provides a more granular view by cross-classifying results by both test length and latent distribution shape.

Several patterns emerge from Figure E.1. First, EQC achieves negligible deviation across all 12 cells of the design, confirming that deterministic root-finding on a fixed quadrature yields numerically exact calibration regardless of test length or latent shape. Second, SAC dispersion decreases systematically with test length: the interquartile range for $I = 60$ is visibly narrower than for $I = 15$. This pattern reflects the reduced Monte Carlo variability of reliability estimators when test information is aggregated over more items. Third, heavy-tailed distributions exhibit somewhat larger SAC outliers, particularly for short tests. This is consistent with the theoretical expectation that heavy-tailed G allocates more probability mass to trait regions where the test information function $\mathcal{J}(\theta; c)$ is low and variable, increasing the variance of the Monte Carlo reliability estimator.

Importantly, median deviations remain close to zero across all cells, indicating that the calibration procedure is approximately unbiased even under substantial departures from normality. The ± 0.02 tolerance bands capture the large majority of SAC conditions in all cells, with coverage improving as test length increases.

E.2. **Finite-Sample Replication Variability.** Section 4.1 noted that $K = 2{,}000$ replicated response datasets were generated per condition to quantify how much the *realized* reliability varies across finite samples, even when the *population-level* design target is held fixed. This distinction is critical for interpreting simulation results: a calibration procedure may achieve the target reliability at the population level, yet individual replications will exhibit sampling variability around that target.

Figure E.2 displays the standard deviation of achieved reliability across the $K = 2{,}000$ replications, stratified by sample size N, test length I, and latent distribution shape.

The results in Figure E.2 reveal three systematic patterns. First, replication variability decreases monotonically with sample size: the median SD drops from approximately 0.03–0.05 at $N = 100$ to approximately 0.01 at $N = 2{,}000$. This $O(N^{-1/2})$ decay is consistent with standard asymptotic theory for reliability estimators. Second, longer tests exhibit lower variability at each sample size, reflecting the law of large numbers applied to item-level information contributions. Third, heavy-tailed

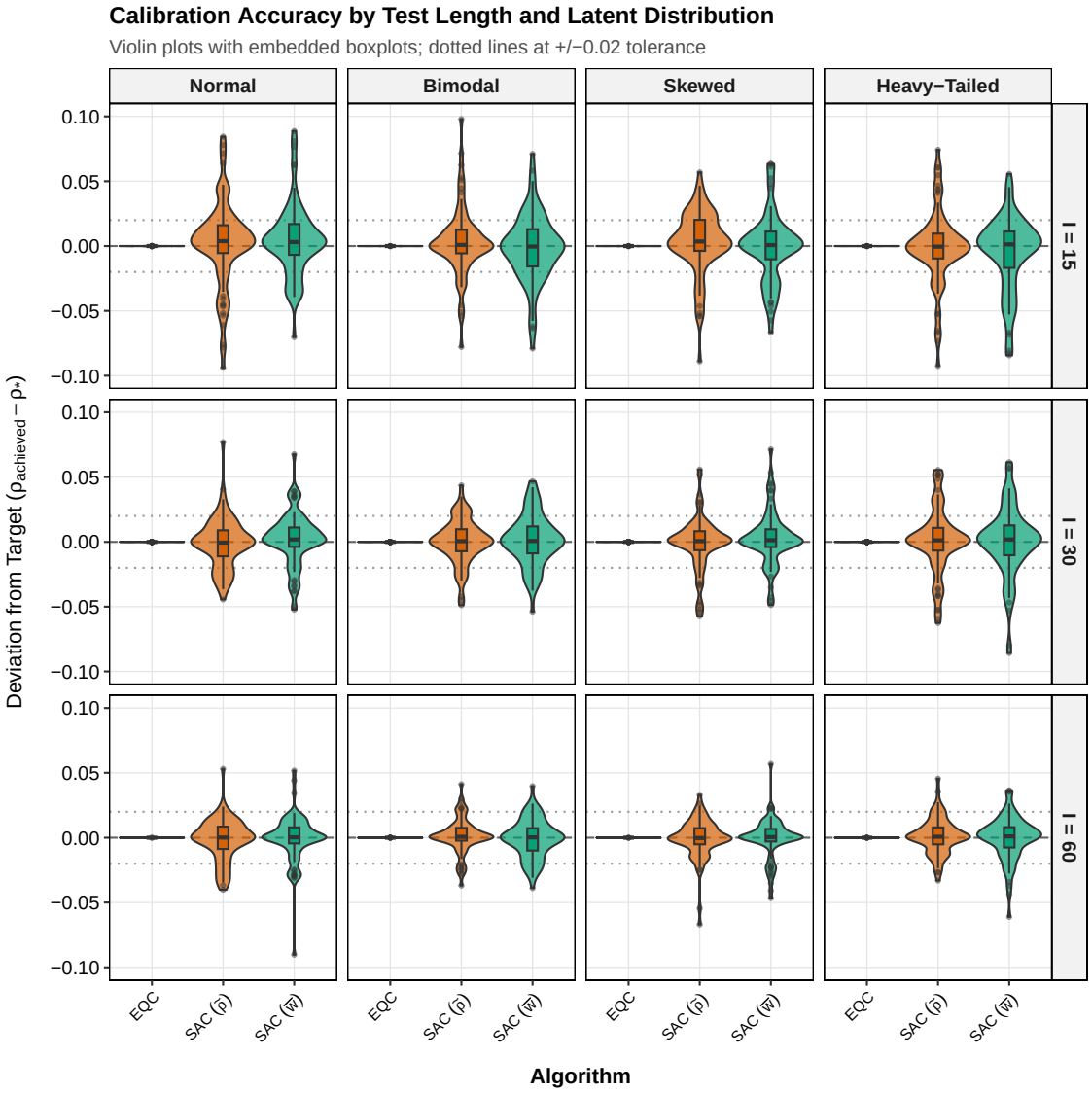

FIGURE E.1. Calibration Accuracy by Test Length and Latent Distribution Shape

Note. Violin plots with embedded boxplots show the distribution of deviations from target reliability ($\Delta = \rho_{\text{achieved}} - \rho^*$) for each algorithm. Rows correspond to test lengths ($I = 15, 30, 60$); columns correspond to latent distribution shapes. The horizontal dashed line indicates $\Delta = 0$ (perfect calibration); dotted lines indicate ± 0.02 tolerance bands. EQC achieves effectively zero deviation across all cells. SAC deviations are centered near zero but exhibit greater dispersion for short tests and non-normal distributions.

distributions exhibit systematically higher variability than normal or bimodal distributions, even after controlling for N and I. This elevation reflects the increased

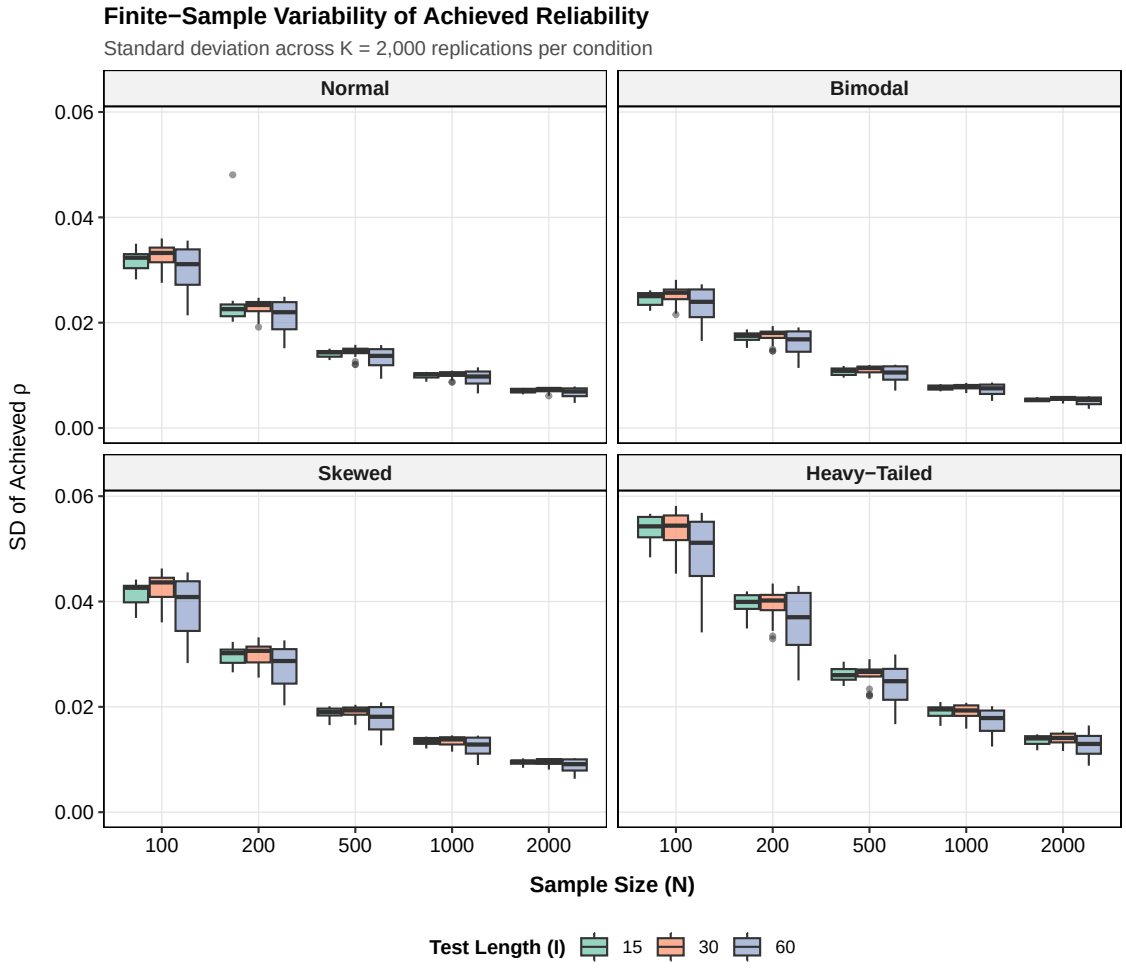

FIGURE E.2. Finite-Sample Variability of Achieved Reliability

Note. Boxplots show the standard deviation of achieved reliability ($\hat{\rho}$) computed across $K = 2{,}000$ replications per condition. Panels correspond to latent distribution shapes; colors indicate test length (I). Within each panel, boxplots are grouped by sample size ($N = 100, 200, 500, 1000, 2000$). Variability decreases with both N and I, and is elevated for heavy-tailed distributions.

probability of extreme θ values where test information is low, which inflates the variance of person-level information contributions.

These findings have practical implications for simulation study design. When sample sizes are small (e.g., $N \leq 200$), researchers should expect substantial replication-to-replication variability in realized reliability, even when calibration is exact at the population level. Reporting the design-level target ρ^* alongside the empirical mean and standard deviation of realized reliability across replications provides a more complete picture of simulation conditions.

E.3. **Calibration Accuracy by Sample Size.** While the main text focused on aggregate calibration accuracy (Table 2), Figure E.3 examines whether calibration error varies systematically with the sample size N used for generating response data.

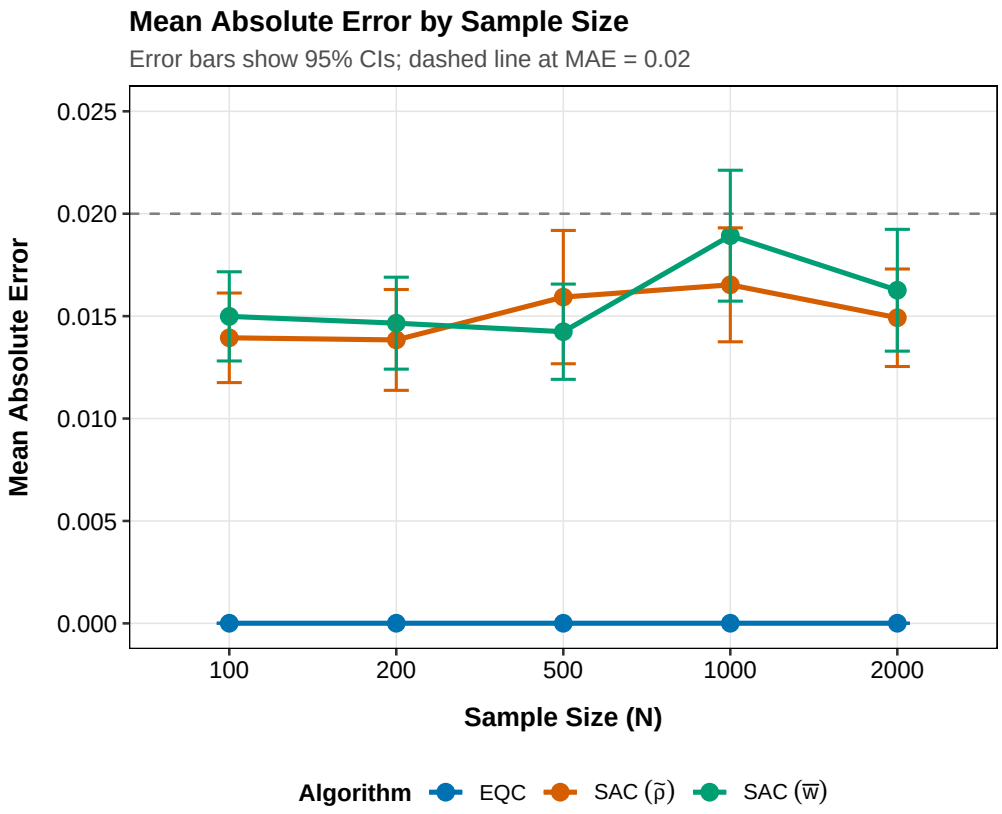

FIGURE E.3. Mean Absolute Error by Sample Size

Note. Points show the mean absolute error (MAE) of achieved reliability relative to the target, averaged across all conditions at each sample size. Error bars indicate 95% confidence intervals for the mean. The horizontal dashed line indicates MAE = 0.02. EQC achieves MAE ≈ 0 regardless of sample size. SAC MAE is approximately constant across sample sizes, indicating that calibration accuracy is determined by the stochastic approximation procedure rather than by the size of the generated datasets.

A key observation from Figure E.3 is that SAC calibration accuracy is essentially invariant to sample size N. This is expected because calibration targets the *population-level* reliability functional $\rho(c)$, which is defined as an expectation over the latent distribution G and does not depend on the sample size used for subsequent data generation. The Monte Carlo draws used within SAC are governed by the stochastic approximation configuration (number of iterations, draws per iteration), not by N.

Consequently, SAC achieves similar MAE (≈ 0.015) whether the generated datasets will contain 100 or 2,000 persons.

This invariance has a practical implication: researchers need not adjust calibration settings based on the intended sample size of generated data. A single calibrated configuration (Ψ, c^*, G) can be used to generate datasets of varying sizes without recalibration.

E.4. **Calibration Success Rate by Target Level and Test Length.** The main text (Table 2) reported that 73.0% of SAC ($\tilde{\rho}$) conditions achieved reliability within ± 0.02 of the target. Figure E.4 disaggregates this success rate by target reliability level ρ^* and test length I, revealing how calibration difficulty varies across the design space.

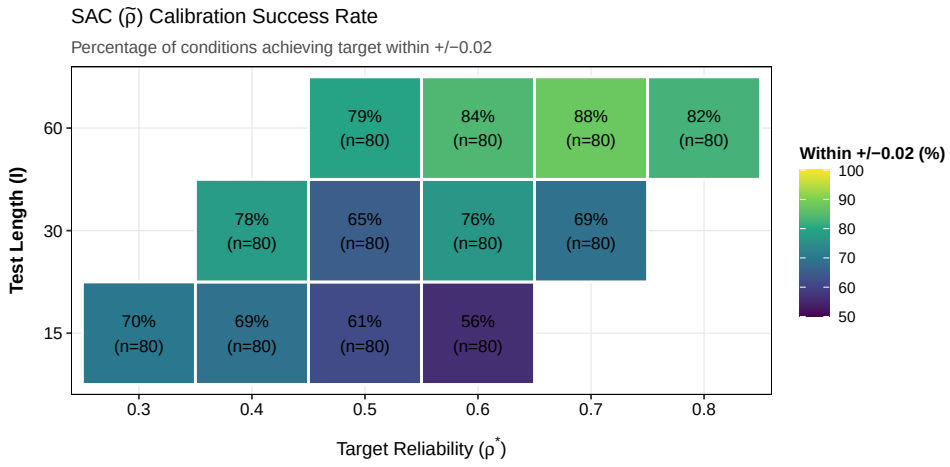

FIGURE E.4. SAC ($\tilde{\rho}$) Calibration Success Rate by Target Reliability and Test Length

Note. Each cell shows the percentage of conditions (out of $n = 80$ per cell) that achieved reliability within ± 0.02 of the target. Darker shading indicates lower success rates. Empty cells correspond to target–test-length combinations excluded by the adaptive target scheme (Section 4.1). Success rates are highest for intermediate targets and longer tests; they are lowest for the most demanding targets at each test length (e.g., $\rho^* = 0.60$ for $I = 15$).

The heatmap in Figure E.4 reveals a consistent pattern: calibration success rates are lower at the boundaries of the feasible target range for each test length. For $I = 15$, the success rate drops to 56% at $\rho^* = 0.60$, which approaches the upper bound of achievable reliability for short tests (Appendix B). For $I = 60$, success rates are uniformly high (79–88%) across all target levels. This pattern is consistent with the feasibility analysis in Section 2.4 and Appendix B: near the boundaries of

the achievable reliability set $\mathcal{R}([c_L, c_U])$, the inverse mapping from ρ^* to c^* becomes increasingly sensitive to Monte Carlo error, leading to larger calibration deviations.

The adaptive target scheme employed in the validation study (Section 4.1) was designed to avoid infeasible targets. Nevertheless, targets near the upper boundary of each test length's feasible range remain more challenging to calibrate precisely. Practitioners seeking high-precision calibration should either (i) use longer tests, (ii) avoid targets near the feasibility boundary, or (iii) increase SAC iterations to reduce stochastic approximation error.

E.5. **Summary.** The extended validation results in this appendix support three conclusions that complement the main text findings:

(1) **Robustness across design factors.** Calibration accuracy is stable across latent distribution shapes, IRT models, and item sources. While heavy-tailed distributions and short tests exhibit somewhat greater SAC variability, the calibration procedure remains approximately unbiased across all 960 conditions.

(2) **Finite-sample variability.** Even with exact population-level calibration, realized reliability in individual replications exhibits sampling variability that decreases with $\sqrt{N}$ and $\sqrt{I}$. Researchers should report both the design target ρ^* and the empirical distribution of realized reliability across replications.

(3) **Boundary sensitivity.** Calibration success rates are lower near the boundaries of the feasible reliability range, where the inverse mapping from ρ^* to c^* becomes increasingly sensitive to Monte Carlo error.

Together with the main text results, these findings validate the reliability-targeted simulation framework as a practical and theoretically grounded approach to IRT data generation.

Appendix F. Software Implementation and Reproducibility

This appendix provides practical guidance for implementing reliability-targeted IRT simulation using the `IRTsimrel` R package. The package is available at https://joonho112.github.io/IRTsimrel/ and implements all methods described in the main text.

F.1. **Package Overview.** The `IRTsimrel` package is designed to make reliability targeting a drop-in component of standard IRT simulation workflows. The implementation separates *structural generators* from *scale calibration*, reflecting the theoretical framework developed in Section 2: realistic item and population features are generated first, and then measurement precision is tuned via global discrimination scaling.

Table F.1 maps the main conceptual components of the paper to the corresponding package functions.

TABLE F.1. Mapping Between Paper Concepts and IRTsimrel Functions

Paper Concept	Function	Description
Latent distribution G	`sim_latentG()`	Generates std. abilities
Item parameters Ψ	`sim_item_params()`	Generates β, λ
EQC algorithm	`eqc_calibrate()`	Deterministic calibration
SAC algorithm	`sac_calibrate()`	Stochastic calibration
Algorithm comparison	`compare_eqc_spc()`	Compares calibrations
Response generation	`simulate_response_data()`	Produces response matrices
External validation	`compute_reliability_tam()`	WLE/EAP reliability
Distribution comparison	`compare_shapes()`	Visualizes latent shapes

Note. See Section 2.1 for latent distributions, Section 3.1 for item parameters, and Sections 3.2 and 3.3 for calibration algorithms. The function `spc_calibrate()` is an alias retained for backward compatibility.

F.2. **Recommended Workflow.** The standard workflow for reliability-targeted simulation consists of four stages: (1) specify structural configuration, (2) calibrate to target reliability, (3) generate response data, and (4) optionally validate. The following code block illustrates an end-to-end example targeting $\rho^* = 0.75$ under a Rasch model with bimodal latent distribution and IRW-based item difficulties.

```
library(IRTsimrel)

# ------------------------------------------------------
# Stage 1: Specify structural configuration
# ------------------------------------------------------
# Latent distribution: bimodal with mode separation delta = 0.8
# Item source: empirical difficulties from Item Response Warehouse
# Model: Rasch (discriminations fixed at 1)
# Test length: 30 items
```

```
# ------------------------------------------------------
# Stage 2: Calibrate using EQC
# ------------------------------------------------------
eqc_result <- eqc_calibrate(
  target_rho      = 0.75,
  n_items         = 30,
  model           = "rasch",
  latent_shape    = "bimodal",
  latent_params   = list(shape_params = list(delta = 0.8)),
  item_source     = "irw",
  reliability_metric = "info",
  M               = 20000,
  c_bounds        = c(0.1, 10),
  seed            = 42
)

print(eqc_result)
#> ========================================================
#>   Empirical Quadrature Calibration (EQC) Results
#> ========================================================
#>
#> Calibration Summary:
#>   Model                        : RASCH
#>   Target reliability (rho*)    : 0.7500
#>   Achieved reliability         : 0.7500
#>   Absolute error               : 2.81e-06
#>   Scaling factor (c*)          : 0.6927
#>
#> Design Parameters:
#>   Number of items (I)          : 30
#>   Quadrature points (M)        : 20000
#>   Reliability metric           : Average-information (tilde)
#>   Latent variance              : 0.9999
#>
```

```
#> Convergence:
#>   Root status                  : uniroot_success
#>   Search bracket               : [0.100, 10.000]
#>   Bracket reliabilities        : [0.0695, 0.9883]
```

The output shows that EQC achieves the target reliability of 0.75 with an absolute error of 2.81×10^{-6}, confirming essentially exact calibration. The calibrated scaling factor is $c^* = 0.693$, which scales down the baseline discriminations to achieve the specified reliability level. The bracket reliabilities $[0.070, 0.988]$ confirm that the target lies well within the feasible range for this configuration.

```
# ------------------------------------------------------
# Stage 3: Generate response data
# ------------------------------------------------------
sim_data <- simulate_response_data(
  eqc_result   = eqc_result,
  n_persons    = 1000,
  latent_shape = "bimodal",
  latent_params = list(shape_params = list(delta = 0.8)),
  seed         = 123
)

dim(sim_data$response_matrix)
#> [1] 1000   30

# ------------------------------------------------------
# Stage 4 (Optional): Validate with TAM
# ------------------------------------------------------
tam_rel <- compute_reliability_tam(
  resp  = sim_data$response_matrix,
  model = "rasch"
)

cat(sprintf("Target:   %.3f\n", eqc_result$target_rho))
cat(sprintf("EAP rel.: %.3f\n", tam_rel$rel_eap))
cat(sprintf("WLE rel.: %.3f\n", tam_rel$rel_wle))
#> Target:   0.750
```

```
#> EAP rel.: 0.753
#> WLE rel.: 0.737
```

External validation via TAM confirms that the realized reliabilities (EAP = 0.753, WLE = 0.737) closely match the design target of 0.75, with the small discrepancies reflecting finite-sample variability and the distinction between population-level and estimator-specific reliability definitions.

For studies requiring independent stochastic validation or direct targeting of MSEM-based reliability $\bar{w}$, SAC can be run after EQC using a warm start:

```
# SAC validation with EQC warm start
sac_result <- sac_calibrate(
  target_rho = 0.75,
  n_items    = 30,
  model      = "rasch",
  latent_shape = "bimodal",
  latent_params = list(shape_params = list(delta = 0.8)),
  item_source = "irw",
  c_init     = eqc_result,
  n_iter     = 1000,
  M_per_iter = 2000,
  seed       = 456
)

compare_eqc_spc(eqc_result, sac_result)
#> ========================================================
#>   EQC vs SPC Comparison
#> ========================================================
#>
#>   Target reliability  : 0.7500
#>   EQC c*              : 0.692742
#>   SPC c*              : 0.722159
#>   Absolute difference : 0.029417
#>   Percent difference  : 4.25%
#>   Agreement (< 5%)    : YES
```

The comparison shows that EQC and SAC yield calibrated scales that agree within 4.25%, well below the 5% threshold for practical equivalence. This independent stochastic validation confirms the accuracy of the deterministic EQC solution.

F.3. **Function Reference Summary. Latent distribution generation.** The `sim_latentG()` function generates abilities from 12 built-in shapes (normal, bimodal, trimodal, skewed, heavy-tailed, uniform, floor/ceiling effects, and custom mixtures), all pre-standardized to $\mathbb{E}[\theta] = 0$ and $\mathrm{Var}(\theta) = 1$ before any location-scale transformation. The `compare_shapes()` function provides side-by-side visualization of multiple shapes for design exploration.

Item parameter generation. The `sim_item_params()` function supports parametric ($\beta \sim N(0, 1)$) and empirical (IRW) difficulty sources. For the 2PL model, discriminations are generated with a target Spearman correlation to difficulties (default $\rho = -0.3$) using the Gaussian copula method described in Appendix D, which preserves exact marginals while achieving the target dependence structure.

Calibration. Both `eqc_calibrate()` and `sac_calibrate()` return objects containing the calibrated scale c^*, achieved reliability, quadrature/iteration diagnostics, and the full item parameter set. The `reliability_metric` argument selects between average-information reliability (`"info"` or `"tilde"`) and MSEM-based reliability (`"msem"` or `"bar"`).

Response generation. The `simulate_response_data()` function accepts either an `eqc_result` or `sac_result` object and generates response matrices under the specified IRT model using the calibrated item parameters. Multiple independent datasets can be generated by varying the seed.

F.4. **Reproducibility.** For exact reproducibility, all stochastic functions in `IRTsimrel` accept a `seed` argument. A complete reproducibility record should include:

(1) **Random seeds** for each stage (calibration, data generation)
(2) **Package version** (`packageVersion("IRTsimrel")`)
(3) **R version** and platform (`sessionInfo()`)
(4) **Key parameter settings** (target reliability, test length, latent shape, item source)

The validation study in Section 4 used the following configuration:

```
# Validation study parameters
M          <- 20000   # Quadrature size for EQC
n_iter     <- 1000    # SAC iterations
```

```
M_per_iter <- 2000    # MC draws per SAC iteration
K          <- 2000    # Replications per condition
c_bounds   <- c(0.1, 10)
```

Complete replication scripts are available at the GitHub repository: https://github.com/joonho112/reliability-targeted-irt-simulation.