

The Beta-Bernstein Bridge: Closed-Form Bayesian Measurement Beyond Monotone Item Response Functions

Yumou Wei

Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15213, USA
yumouw@andrew.cmu.edu

Abstract

Item response theory forces a choice between flexible curves and tractable inference. We bridge them by representing item response functions as Bernstein polynomials. With a natural conjugate Beta prior, this gives a closed-form ability posterior as a Beta mixture. It reproduces IRT and recovers non-monotone curves with bimodal posteriors.

1 Introduction

Item response theory (IRT) measures students' abilities from their responses to assessment items using the item-response function (IRF), which models the probability of a correct response as a function of ability and item parameters. Classic 1PL or 2PL models assume a logistic IRF, but real response data often deviate from these idealized curves (Loken and Rulison, 2010; Culpepper, 2016), which calls for more flexible IRFs (Samejima, 2000; Lee and Bolt, 2018; Liao and Bolt, 2021).

At the same time, accurate and fair measurement requires principled Bayesian inference. A closed-form posterior ability provides the standard error and credible interval of an estimate, and an exact marginal likelihood compares competing models. The chronic difficulty is that flexible IRFs make the marginal likelihood intractable (Bock and Aitkin, 1981), while closed-form Bayesian inference exists only for specific parametric IRFs (Polson et al., 2013; Durante, 2019) with limited flexibility. Flexibility and tractability pull in opposite directions.

A desirable flexibility is relaxing the assumption that IRFs must be monotone (increasing). Nonparametric IRT routinely tests this assumption (Junker and Sijtsma, 2000), and unfolding models of attitude measurement have non-monotone IRFs built in (Roberts et al., 2000). A setting where the monotonicity assumption may fail in *ability* measurement is **AI-integrated assessment**, in which students consult an AI assistant while responding.

Field experiments find that AI assistance helps lower-ability users the most (Noy and Zhang, 2023; Brynjolfsson et al., 2025), that human-AI teams do not simply inherit the AI's accuracy (Vaccaro et al., 2024), and that reliance on AI falls with the participant's own ability in some released human-AI assessment data (Section 4.2.3). If a competent assistant is used most by those who need it most, a low-ability student may answer a hard item correctly by following the AI while a mid-ability student who overrides it does not (Figure 1, left). A correct response is therefore genuinely ambiguous, and only a flexible, exact posterior can express this ambiguity with two modes, which a Gaussian approximation can never show (Figure 1, right).

We propose the **Beta-Bernstein Bridge**, which connects any continuous IRF on $[0, 1]$ to a conjugate Beta prior. Representing IRFs as Bernstein polynomials lets us exploit the conjugacy of the Beta distribution with the Binomial likelihood, so that the exact posterior ability is a finite mixture of Beta distributions for any IRF, monotone or not.

We evaluated the bridge on six real student datasets and one simulated dataset with AI-mediated responses, giving affirmative answers to two research questions:

RQ-1: Do bridged models reproduce the predictions of their standard counterparts? (Section 4.1)

RQ-2: Can the bridge represent non-monotone IRFs? (Section 4.2)

Based on our results, we developed some guidelines for practitioners to decide whether the bridge is worthwhile, including its computational cost, degree selection and parameter scaling (Section 5).

2 Related Work

Neither flexible IRFs nor closed-form ability posteriors are new, but no prior method delivers exact Bayesian inference under an arbitrary IRF.

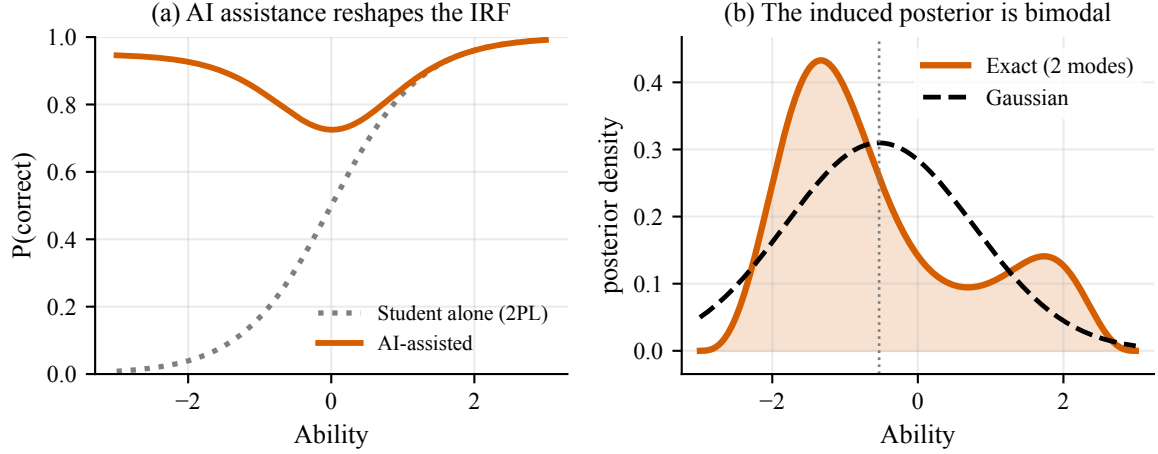


Figure 1: AI-integrated assessment can produce atypical item-response curves that deviate from the traditional monotonic shape (left; the generative model used in Section 4.2.1). Exact Bayesian inference produces a *bimodal* posterior over ability (right), which reflects the uncertainty about whether a correct response is due to the student’s ability or the AI’s assistance. An approximate Gaussian posterior would fail to capture the bimodality.

Flexible IRFs preclude closed-form posteriors. Kernel-smoothed item characteristic curves (Ramsay, 1991) and monotone polynomials (Falk and Cai, 2016) are flexible, but have no closed-form posterior over ability. Bayesian IRT models place priors on the IRF through infinite mixtures (Karabatsos, 2017) or Gaussian processes (Duck-Mayr et al., 2020), or use variational inference (Wu et al., 2020), but all return approximate or sampled posteriors and no exact marginal likelihood. We instead choose an IRF representation with a natural conjugate prior so that any IRF expressed as such inherits a closed-form posterior.

Exact inference exists only for specific IRFs. Durante (2019) shows that the probit likelihood is conjugate to a Gaussian prior, with a unified skew-normal posterior, but the result is probit-specific and framed for regression coefficients. Polson et al. (2013) make the logistic likelihood conditionally Gaussian by Pólya-Gamma augmentation, but the method is logistic-only and still needs Gibbs sampling. The closest IRT construction (Liu and Wang, 2020) places a monotone nonparametric prior on a *dynamic* ability trajectory under a fixed logistic IRF, solved by MCMC. While previous methods retain tractability by fixing the IRF, our method extends to any bounded IRF representable as a Bernstein polynomial, whether it is monotone or not.

Bernstein polynomials have been used, but not for conjugacy. They served as Bayesian nonparametric priors for densities and shape-constrained regression (Petrone, 1999; Curtis and Ghosh, 2011; Choi et al., 2016), and in integrated

(I-spline) form imposed monotonicity through non-negative coefficients (Ramsay, 1988). Previous work used them to *estimate a function* by posterior sampling; we exploited their algebraic properties to *represent a function* so that the posterior is available in closed form, which is a novel use in IRT.

3 The Beta-Bernstein Bridge

3.1 Problem Statement

Consider measuring a student’s latent ability θ_i from their responses to a set of assessment items. In IRT, each item j has an **item-response function (IRF)** $P_j(\theta_i; \mathbf{w}_j)$ that relates the probability of a correct response to θ_i given the item parameters $\mathbf{w}_j$. The responses form a vector of S_i dichotomous outcomes $\mathbf{y}_i = (y_{i1}, \dots, y_{iS_i})$, where each entry is 1 if the item is answered correctly and 0 otherwise.

We seek to perform **Bayesian measurement** by placing a prior $p(\theta_i)$ on the latent ability and computing the posterior $p(\theta_i | \mathbf{y}_i)$ via the Bayes rule. If the responses are conditionally independent given θ_i , the likelihood is a product of the IRFs:

$$p(\mathbf{y}_i | \theta_i) = \prod_{j=1}^{S_i} P_j(\theta_i)^{y_{ij}} (1 - P_j(\theta_i))^{1-y_{ij}}. \quad (1)$$

Calculating the posterior in closed form requires first computing the exact marginal likelihood $p(\mathbf{y}_i)$, which results from integrating the likelihood against the prior over all values of θ_i :

$$p(\mathbf{y}_i) = \int p(\theta_i) \prod_{j=1}^{S_i} P_j(\theta_i)^{y_{ij}} (1 - P_j(\theta_i))^{1-y_{ij}} d\theta_i. \quad (2)$$

This integral is generally intractable, and a closed-form solution is only known for a few specific IRFs. Our goal is to find a novel IRF representation such that the integral can be solved in closed form for any IRF bounded on $[0, 1]$.

3.2 Flexible Representation of IRFs

Canonical IRT models place the ability θ on the real line, with a practical range of $[-3, 3]$ under a standard Normal prior. We instead place θ on the unit interval $[0, 1]$ under a **Beta** prior $\text{Beta}(\alpha, \beta)$ and represent each IRF as a **Bernstein polynomial** (Bernstein, 1912), whose flexibility and close relationship with the Beta distribution yield a closed-form posterior for any bounded IRF.

3.2.1 Bernstein Polynomials

A Bernstein polynomial of degree n is a linear combination of $n + 1$ *basis functions*:

$$b_{k,n}(\theta) = \binom{n}{k} \theta^k (1 - \theta)^{n-k}, \quad k = 0, \dots, n. \quad (3)$$

The basis functions are precisely the Binomial probability mass function $\text{Bin}(k; n, \theta)$ viewed as a function of θ , with two key properties. First, they are non-negative and sum to one for all $\theta \in [0, 1]$, so any linear combination of them with coefficients in $[0, 1]$ is a valid probability function. Second, they are conjugate to the Beta distribution, following directly from the Beta-Binomial conjugacy.

We represent each item’s IRF as a Bernstein polynomial with coefficients $\{w_k^{(j)}\}_{k=0}^n$:

$$P_j(\theta) = \sum_{k=0}^n w_k^{(j)} b_{k,n}(\theta). \quad (4)$$

The coefficients $w_k^{(j)}$ are the item parameters that determine the shape of the IRF. They are constrained to $[0, 1]$ so that $P_j(\theta)$ is a valid probability; to enforce monotonicity, they can further be constrained to be non-decreasing, $w_0^{(j)} \leq \dots \leq w_n^{(j)}$.

Intuitively, the ability continuum is divided into $n + 1$ “cognitive states” at the nodes k/n , where $w_k^{(j)}$ is the probability of giving a correct response to item j in the k -th state, $b_{k,n}(\theta)$ the probability of being in that state given θ , and the degree n sets the “resolution” of the states.

3.2.2 Universal Approximation

A classical result from the Weierstrass approximation theorem (Bernstein, 1912; Lorentz, 1953) specifies exactly how to choose the coefficients to

represent any continuous IRF on $[0, 1]$: for any continuous g on $[0, 1]$, the Bernstein polynomial with coefficients $w_k = g(k/n)$ converges uniformly to g as $n \rightarrow \infty$:

$$\lim_{n \rightarrow \infty} \sup_{\theta \in [0,1]} \left| g(\theta) - \sum_{k=0}^n g(k/n) b_{k,n}(\theta) \right| = 0. \quad (5)$$

Any continuous IRF can thus be bridged into the Beta-Bernstein family, to arbitrary precision, by sampling it at the nodes k/n . The logistic and normal-ogive curves, for instance, are defined on the unbounded ability scale $\mathbb{R}$, so we first map $\theta \in [0, 1]$ to a wider interval $[-h, +h]$ with the link function $\eta = 2h\theta - h$ and sample the curve at the mapped nodes: $w_k = \sigma(\eta(k/n))$ (logistic) or $w_k = \Phi(\eta(k/n))$ (normal-ogive).

Alternatively, the coefficients can be learned from data without any parametric form; the theorem still guarantees that the learned polynomial can represent any continuous IRF if n is large enough. To keep the number of item parameters from growing with n , we can learn a *coefficient-generating function* with far fewer parameters—for example, a monotone I-spline (Ramsay, 1988) with $m \ll n$ parameters sampled at the nodes k/n —which decouples the degree n from the number of item parameters m . For maximum flexibility, the per-item coefficients can also be optimized directly, optionally under a monotonicity constraint, giving a fully nonparametric IRF.

3.3 Beta-Bernstein Conjugacy

The item parameters, whether generator parameters or the coefficients themselves, are fitted by maximizing the marginal likelihood $p(\mathbf{y}_i)$, whose exact form is needed for both estimation and posterior inference. Define the effective weights for item j as $\tilde{w}_k^{(j)} = w_k^{(j)} y_{ij} + (1 - w_k^{(j)})(1 - y_{ij})$, contingent on the response y_{ij} . The marginal likelihood is

$$p(\mathbf{y}_i) = \int_0^1 \text{Beta}(\theta_i; \alpha, \beta) \prod_j^{S_i} \left[\sum_{k=0}^n \tilde{w}_k^{(j)} b_{k,n}(\theta_i) \right] d\theta_i. \quad (6)$$

The integrand is a product of S_i Bernstein polynomials, and two elementary properties of Bernstein polynomials give the marginal likelihood and the posterior in closed form.

First, a product of Bernstein polynomials is again a Bernstein polynomial. Multiplying two Bernstein polynomials boils down to the pairwise products of their basis functions, which result in a

Dataset	Students	Items	% Correct	Brief Description
art (McCarron and Kuperman, 2021)	1,402	50	27.08%	Author recognition
spelling (Eskenazi et al., 2022)	673	109	56.32%	Elementary spelling
malawi_math (Berry et al., 2022)	6,818	10	43.97%	Math achievement
online_rct (Gilbert et al., 2023)	7,797	30	53.72%	Reading comprehension
content_lit (Kim et al., 2023)	2,174	20	54.87%	Reading comprehension
vocab (Kim et al., 2021)	2,588	24	44.45%	Vocabulary knowledge

Table 1: Summary of the six datasets used in the experiments, including the number of students, items, overall percentage of correct responses, and a brief description of the assessment context.

(scaled) basis function of a higher degree:

$$b_{s,m}(\theta) \cdot b_{t,n}(\theta) = \frac{\binom{m}{s} \binom{n}{t}}{\binom{m+n}{s+t}} b_{s+t, m+n}(\theta). \quad (7)$$

Hence the product of S_i Bernstein polynomials resolves into a single Bernstein polynomial of degree $D_i = S_i \cdot n$ with the coefficients $\{C_\kappa\}_{\kappa=0}^{D_i}$ that depend on all item weights and responses:

$$\prod_j \left[\sum_{k=0}^n \tilde{w}_k^{(j)} b_{k,n}(\theta) \right] = \sum_{\kappa=0}^{D_i} C_\kappa b_{\kappa, D_i}(\theta). \quad (8)$$

Second, a Bernstein basis function is conjugate to a Beta prior. Each basis function is a Binomial likelihood, so integrating it against a Beta prior yields the Beta-Binomial marginal likelihood:

$$\int_0^1 \text{Beta}(\theta; \alpha, \beta) b_{\kappa, D_i}(\theta) d\theta = \binom{D_i}{\kappa} \frac{B(\alpha + \kappa, \beta + D_i - \kappa)}{B(\alpha, \beta)}. \quad (9)$$

Since Equation 8 is a sum of $D_i + 1$ basis functions weighted by C_κ , the marginal likelihood simply sums their contributions:

$$p(\mathbf{y}_i) = \sum_{\kappa=0}^{D_i} C_\kappa \binom{D_i}{\kappa} \frac{B(\alpha + \kappa, \beta + D_i - \kappa)}{B(\alpha, \beta)} \quad (10)$$

and the posterior is a $D_i + 1$ mixture of the corresponding Beta-Binomial posteriors:

$$p(\theta_i | \mathbf{y}_i) = \sum_{\kappa=0}^{D_i} \pi_\kappa \text{Beta}(\theta_i; \alpha + \kappa, \beta + D_i - \kappa),$$

with $\pi_\kappa \propto C_\kappa \binom{D_i}{\kappa} B(\alpha + \kappa, \beta + D_i - \kappa).$

(11)

Regardless of IRF, the posterior is a finite mixture of Beta distributions, and the marginal likelihood is a finite sum of Beta functions. This is the **Beta-Bernstein Bridge** connecting any continuous

S	n	Comps.	Exact	Quad.	Direct
10	10	101	4.7	3.5	1.2
10	20	201	8.8	5.0	1.2
30	10	301	23.1	7.9	2.5
30	20	601	33.1	11.0	2.5
50	10	501	40.6	12.0	5.3
50	20	1,001	70.5	17.0	5.3
109	10	1,091	131.6	23.6	6.9
109	20	2,181	517.0	34.2	6.9

Table 2: Cost of one marginal-likelihood evaluation for 256 students (milliseconds, median of 10 runs; PyTorch on an Apple M2 Pro CPU) as a function of test length S and degree n , for the exact method, the Gauss-Jacobi quadrature form of the same closed form ($Q = 80$ nodes), and a direct Gauss-Hermite 2PL ($Q = 80$). “Comps.” is the number of components $Sn + 1$ of the exact posterior. A training step adds a backward pass of comparable cost.

IRF on $[0, 1]$ with a conjugate Beta prior, which enables exact Bayesian inference for any IRF representable as a Bernstein polynomial.

Equation 8 multiplies S_i polynomials of degree n , which costs $O(D_i^2)$ per student and yields a posterior with $D_i + 1$ components. The same integrals can also be evaluated by the Gauss-Jacobi quadrature with Q nodes at a cost of $O(D_i Q)$; because the integrand is a polynomial of degree D_i weighted by a Beta density, the quadrature is *exact* if $Q \geq \lceil (D_i + 1)/2 \rceil$ for all i . Therefore, the exact and quadrature evaluations are interchangeable. Table 2 compares both with a standard Gauss-Hermite 2PL: the quadrature costs 2–5 $\times$ and the exact costs 4–75 $\times$ higher than a direct evaluation—about two milliseconds per student at worst.

4 Experiments

We evaluated the Beta-Bernstein Bridge around the two research questions introduced in Section 1.¹

¹Code, benchmark configurations and simulation scripts are available at <https://github.com/weiyumou/BetaBern>.

IRF family	Direct	Bridged ($n = 10$)	Bridged ($n = 20$)	Equiv.	# Params
Logistic (2PL)	0.5064	0.5073	0.5069	6/6	83
Normal-ogive (2PNO)	0.5072	0.5074	0.5072	6/6	83
I-spline	0.5051	0.5066	0.5054	6/6	116
Free-form	–	0.5061	0.5050	–	114 & 144

Table 3: Held-out NLL (averaged over six datasets; lower is better) for the **direct** (standard IRT on $\mathbb{R}$) and **bridged** (exact) forms at degrees $n = 10$ and $n = 20$; the “Equiv.” column counts datasets where the bridged model is TOST-equivalent to its direct counterpart at the margin of 0.01 NLL, *for both degrees* (per-dataset results in Table 6).

4.1 RQ-1: Do bridged models reproduce the predictions of the standard counterparts?

The first experiment probes **faithfulness**: perfect representation of a given IRF is only guaranteed in the limit of an infinite degree, so we stress-test the bridge across six real student datasets and three IRF families to confirm that bridged models are statistically equivalent to the original models at feasible degrees of $n = 10$ and $n = 20$.

4.1.1 Data and Models

We drew six educational assessment datasets from the Item Response Warehouse (Domingue et al., 2025). Each measures a single construct with one fixed ability per student, and the items are scored dichotomously (Table 1).

Following Section 3.2, we tested three IRF families: two parametric with the two-parameter *logistic* (2PL) and *normal-ogive* (2PNO) curves, and one nonparametric with an *I-spline* generator. Each family was evaluated in two forms: a **direct** form serving as the control, i.e., a standard IRT model under a standard Normal prior, and a **bridged** form with the same IRF represented as an n -th degree Bernstein polynomial under a Beta prior by sampling it at the nodes k/n . The two forms have the same number of parameters, so we expect them to have equivalent predictive power.

We also fit a fully **free-form** model that learns the Bernstein coefficients of a monotone IRF directly from data. It exists only in bridged form and is not expected to be equivalent to any family, but it tests the bridge’s ability to represent a curve with no inductive bias other than monotonicity; we expect it to parallel the direct forms, if not better.

4.1.2 Evaluation Procedures

We fit every model by maximum marginal likelihood (MML): the *bridged* models with the exact closed-form marginal likelihood of Section 3.3 and

the *direct* models with Gauss-Hermite quadrature, the standard method for IRT.

For each dataset, we applied ten-fold cross-validation *within each student*: the response vector y_i was randomly partitioned into ten equal-sized folds, so every student contributes to every training fold and the ability distribution is identical across folds by construction (Appendix A documents fold balance and all training settings). In each iteration, we trained all models on the nine training folds aggregated over students and evaluated them on the held-out fold. We report the average negative log-likelihood (NLL) of the held-out responses under each model’s posterior predictive, given the student’s training responses.

Because we ask whether bridged and direct models are *equivalent*, we test equivalence rather than difference with the **two one-sided tests (TOST)** procedure (Lakens, 2017; Lakens et al., 2018), which declares two models equivalent only when their difference is provably smaller than a pre-set margin (0.01 nats of NLL, a tight margin given the NLLs we observe), with p -values corrected for multiple comparisons by the Benjamini and Hochberg (1995) procedure at a false discovery rate of 0.05.

4.1.3 Results and Discussion

Table 3 summarizes the results. **The bridged models reproduce their standard counterparts.** The bridged models (with exact marginal likelihood) are TOST-equivalent to their direct counterparts **in all six datasets, for all three families, and at both degrees.** Re-expressing a monotone IRF as a Bernstein polynomial does not change the model’s predictions, even at a degree as low as $n = 10$, and the finite-degree approximation gap shrinks as the degree grows (bridged logistic went $0.5073 \rightarrow 0.5069$ as n went $10 \rightarrow 20$), exactly reflecting the uniform convergence of Section 3.2.2.

A learned IRF is also faithful. The free-form model, which learns the Bernstein coefficients di-

rectly, also tracks the direct models closely, with an average NLL of 0.5061 at $n = 10$ and 0.5050 at $n = 20$. The bridge can thus represent flexible IRFs without sacrificing predictive power, and the exact marginal likelihood is a reliable objective for fitting flexible IRFs to student data.

4.2 RQ-2: Can the bridge represent non-monotone IRFs?

The second experiment evaluates **flexibility**, the bridge’s most distinctive capability. The conjugacy of Section 3.3 requires only that the coefficients w_k be bounded in $[0, 1]$, *not monotone*, so the closed-form posterior holds even for an IRF that rises *and* falls, which is impossible for standard monotone models. We test this capability in a simulated scenario of **AI-integrated assessment**, one plausible setting in which non-monotonicity can arise as students get help from an AI assistant.

4.2.1 Data and Models

We simulated a bank of ten items: four *standard* 2PL items and six *AI-mediated* items. For an AI-mediated item j , the IRF at ability η is

$$P_j(\eta) = (1 - r_j(\eta)) \cdot \sigma(a_j(\eta - b_j)) + r_j(\eta) \cdot A_j \quad (12)$$

where σ is the logistic function and a_j, b_j are 2PL item parameters. A_j is the AI’s competence (the probability of a correct response if the student fully relies on AI), and $r_j(\eta) = \sigma(-\eta/s_r)$ is the reliance of a student with ability η on AI, with s_r controlling the strength of the AI effect (smaller s_r , stronger effect). We set $A_j = 0.95$ and $s_r = 0.55$ for a pronounced effect. We ran 20 simulations, each with 2,000 students whose abilities were drawn from a standard Normal prior.

The response thus mixes the student’s own ability and the AI’s assistance, with lower-ability students relying more on AI. The resulting IRF is non-monotone as shown on the right of Figure 2: low-ability students ride the AI to a correct response, mid-ability students who override the AI but lack the skill dip down, and high-ability students recover on their own. A correct response is thus ambiguous between low ability (rode the AI) and high ability (did it alone)—a pattern that monotone models cannot capture. We compared three MML-fitted models: a four-parameter logistic (4PL) model (the strongest monotone parametric baseline), the **monotone** free-form Bernstein model of Section 4.1, and a **flexible** free-form Bernstein model without monotonicity constraints.

4.2.2 Evaluation Procedures

Because the true IRF of every item is known, we evaluate how well each model recovers it using the median root mean squared error (RMSE) between each fitted IRF and the true curve, separately for standard and AI-mediated items. This requires no held-out set, as a spurious wiggle only raises the error against the truth. We also report AIC and BIC to confirm that any advantage is not merely due to the extra parameters, since the flexible model carries four more parameters than the 4PL.

4.2.3 Results and Discussion

Only the flexible model recovers the valley. On standard items all three models recover the true curve with an RMSE of 0.03–0.06, and the flexible model (0.028) slightly outperformed the monotone Bernstein (0.033) and the 4PL (0.056), so relaxing monotonicity does not hurt on monotone items. On AI-mediated items, the flexible model traces the valley within an RMSE of 0.033, whereas the monotone 4PL and Bernstein models are forced to rise through the dip and miss it with roughly a $5\times$ larger median error (Table 4, Figure 2).

The advantage survives a capacity penalty. The flexible model wins on both AIC and BIC despite its four extra parameters over the 4PL model, while the *monotone* Bernstein model, parameter-matched to the flexible model, is the worst on both criteria (Table 4). Extra capacity does not help, unless spent on relaxing the monotonicity constraint.

The closed form yields a non-Gaussian posterior. For a student who answers the AI-mediated items correctly, the exact Beta-mixture posterior over ability is **bimodal** (Figure 1, right): a “low ability, rode the AI” mode and a “high ability, did it alone” mode. A Gaussian summary has a single peak between the two modes and represents neither, so only the closed form captures this ambiguity.

Empirical status. The simulation is a capability demonstration, so it is reasonable to consider whether the same story extends to real data. We analyzed ChatBench (Chang et al., 2025), a released corpus in which crowd workers answered MMLU items alone and then with GPT-4o. Its ingredients are documented: when the AI’s answer differed from a worker’s own, the worker adopted it less often if they had a higher solo accuracy ($q = 0.006$) and higher stated confidence ($q < 10^{-5}$), and adoption was partially calibrated to the AI’s competence (83% when the AI was right and 42% when it was wrong; worker-clustered tests, BH-corrected).

Model	# Params	AIC	BIC	RMSE (standard)	RMSE (AI-mediated)
Monotone 4PL	55	22,565	22,873	0.056	0.164
Monotone Bernstein	59	22,613	22,943	0.033	0.169
Flexible Bernstein	59	22,196	22,526	0.028	0.033

Table 4: IRF-recovery RMSE against the true curve (lower is better) on standard vs. AI-mediated items, along with capacity-penalized AIC/BIC under a pronounced AI effect. All quantities are medians over 20 simulation runs. The monotone Bernstein model is parameter-matched to the flexible model.

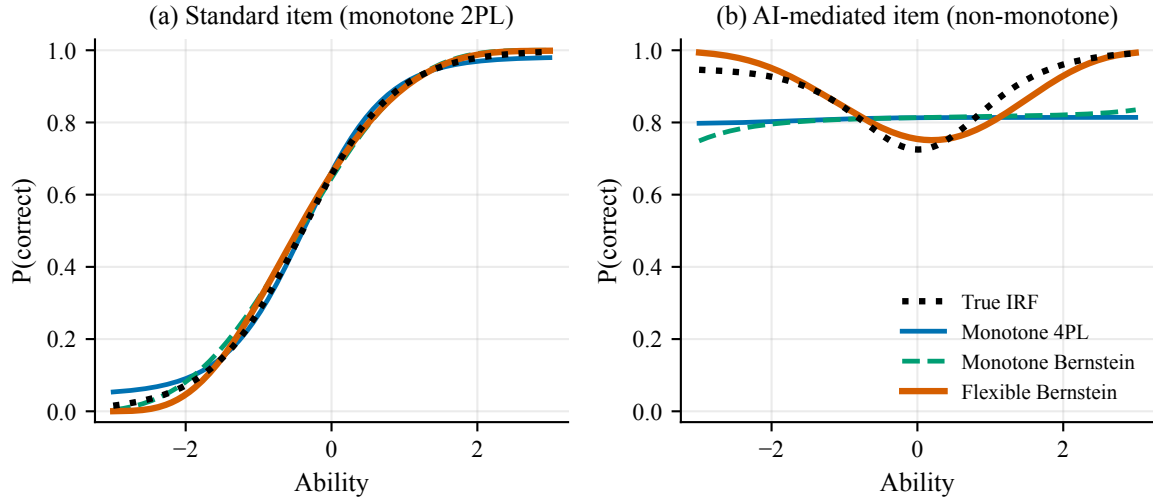


Figure 2: Fitted IRFs on a standard item (left, monotone) and an AI-mediated item (right, a non-monotone valley). The flexible model traces the valley, while the monotone 4PL and Bernstein models are forced to rise through it.

The presence of a valley is suggested but not established: on items where AI outperformed the workers, worker-level AI-assisted accuracy is convex in solo ability with a mid-ability dip of 0.08–0.14, but the contrast against control items does not survive BH correction ($q = 0.055$) and the controls are confounded (one at ceiling, the other with $n = 65$). Whether AI mediation produces population-level valleys in operational assessments therefore remains an open question.

5 General Discussion: When should a practitioner use the bridge?

The bridge is strictly more general than a logistic IRF, but not free. A natural question to ask then is: when to use the bridge? We offer some guidelines based on what we observed during our experiments.

On monotone items, the bridge adds an exact posterior. On the six datasets of Section 4.1, dropping monotonicity moved held-out NLL by between +0.0001 and −0.009 nats (Table 6). The flexibility is neither needed nor harmful on conventional assessments. What the bridge adds there is an exact posterior with analytic mean, variance and

credible intervals, and an exact marginal likelihood for model comparison, at the cost of milliseconds per student (Table 2).

Choose the estimator by test length, the degree by resolution. The exact method works best for short-to-moderate test lengths and sequential scoring, where the posterior after each item is a closed-form Beta mixture; its cost grows quadratically with Sn . The quadrature form grows linearly and works better for longer tests: a few dozen nodes already match the exact method to about 10^{-4} nats on a 109-item test. Degree governs resolution, not capacity: from $n = 10$ to 20 the average held-out NLL moved by at most 0.0012 nats (Table 3), and at $n = 100$ (quadrature form) the residual gap to the direct logistic model was 0.0003 nats, against 0.0009 at $n = 10$. Start at $n = 10$ and raise it only for curves with features narrower than the Bernstein kernel, roughly $1/(2\sqrt{n})$ on the unit scale.

Parameter scaling. The number of item parameters need not grow with n : parametric families keep two parameters per item, an I-spline generator uses $m \ll n$, and even the free-form model uses a rank-2 factorization of its weight matrix (Appendix A), so per-item parameters stay at two while

only a shared $(n + 1)$ -vector grows with degree.

Match the Beta prior to the link function. The bridge samples a curve on $[-h, h]$ via the link function $\eta = 2h\theta - h$. Matching the direct models' normal prior $\mathcal{N}(0, 1)$, i.e., keeping the implied prior on η at unit variance, requires $\alpha = \beta = (h^2 - 1)/2$, so the results in Table 3 use $h = 3$ with a Beta(4, 4) prior. A uniform Beta(1, 1) prior is much wider ($\text{Var}(\eta) = 3$) and under-regularizes, so h and the prior should be chosen together.

6 Conclusion

The Beta-Bernstein Bridge connects any continuous IRF on $[0, 1]$ to a conjugate Beta prior through a shared representation space created by Bernstein polynomials, making exact Bayesian inference available for flexible and even non-monotone IRFs that standard IRT cannot offer. On six real datasets and three IRF families, bridged models reproduce their standard counterparts; in a simulated AI-integrated assessment, the bridge recovers a non-monotone IRF that monotone baselines miss, even under a capacity penalty, and yields a bimodal posterior showing clear uncertainty. Future work will test the bridge on real AI-assistance data and extend it to polytomous items and adaptive testing.

Limitations

The non-monotone demonstration is simulated: the reliance mechanism it assumes is documented in human-AI data, but a population-level valley in an operational assessment has not been established (Section 4.2.3), and relaxing monotonicity did not change held-out predictions on six conventional assessments (Section 5). The formulation covers dichotomous items only. Finally, the bridge inherits IRT's unidimensionality and local independence assumptions; violating either can mimic a non-monotone curve on a monotone item, which our model cannot detect.

References

- Yoav Benjamini and Yoel Hochberg. 1995. [Controlling the false discovery rate: A practical and powerful approach to multiple testing](#). *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.
- Serge Bernstein. 1912. Démonstration du théorème de Weierstrass fondée sur le calcul des probabilités. *Communications de la Société Mathématique de Kharkov*, 13(1):1–2.
- James Berry, Hyuncheol Bryant Kim, and Hyuk Son. 2022. [When student incentives do not work: Evidence from a field experiment in Malawi](#). *Journal of Development Economics*, 158:102893.
- R. Darrell Bock and Murray Aitkin. 1981. [Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm](#). *Psychometrika*, 46(4):443–459.
- Erik Brynjolfsson, Danielle Li, and Lindsey R. Raymond. 2025. [Generative AI at work](#). *The Quarterly Journal of Economics*, 140(2):889–942.
- Serina Chang, Ashton Anderson, and Jake M. Hofman. 2025. [ChatBench: From static benchmarks to human-AI evaluation](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26009–26038, Vienna, Austria. Association for Computational Linguistics.
- Taeryon Choi, Hea-Jung Kim, and Seongil Jo. 2016. [Bayesian variable selection approach to a Bernstein polynomial regression model with stochastic constraints](#). *Journal of Applied Statistics*, 43(15):2751–2771.
- Steven Andrew Culpepper. 2016. [Revisiting the 4-parameter item response model: Bayesian estimation and application](#). *Psychometrika*, 81(4):1142–1163.
- S. McKay Curtis and Sujit K. Ghosh. 2011. [A variable selection approach to monotonic regression with Bernstein polynomials](#). *Journal of Applied Statistics*, 38(5):961–976.
- Benjamin W. Domingue, Mika Braginsky, Lucy Caffrey-Maffei, Joshua B. Gilbert, Klint Kanopka, Radhika Kapoor, Hansol Lee, Yiqing Liu, Savira Nadela, Guanzhong Pan, Lijin Zhang, Susu Zhang, and Michael C. Frank. 2025. [An introduction to the Item Response Warehouse \(IRW\): A resource for enhancing data usage in psychometrics](#). *Behavior Research Methods*, 57:276.
- JBrandon Duck-Mayr, Roman Garnett, and Jacob M. Montgomery. 2020. [GPIRT: A Gaussian process model for item response theory](#). In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *PMLR*, pages 520–529.
- Daniele Durante. 2019. [Conjugate Bayes for probit regression via unified skew-normal distributions](#). *Biometrika*, 106(4):765–779.
- Michael A. Eskenazi, Robert L. Askew, and Jocelyn R. Folk. 2022. [Precision in the measurement of lexical expertise: The selection of optimal items for a spelling assessment](#). *Behavior Research Methods*, 55(2):623–632.
- Carl F. Falk and Li Cai. 2016. [Maximum marginal likelihood estimation of a monotonic polynomial generalized partial credit model with applications to multiple group analysis](#). *Psychometrika*, 81(2):434–460.

- Joshua B. Gilbert, James S. Kim, and Luke W. Miratrix. 2023. [Modeling item-level heterogeneous treatment effects with the explanatory item response model: Leveraging large-scale online assessments to pinpoint the impact of educational interventions](#). *Journal of Educational and Behavioral Statistics*, 48(6):889–913.
- Brian W. Junker and Klaas Sijtsma. 2000. [Latent and manifest monotonicity in item response models](#). *Applied Psychological Measurement*, 24(1):65–81.
- George Karabatsos. 2017. [A menu-driven software package of Bayesian nonparametric \(and parametric\) mixed models for regression analysis and density estimation](#). *Behavior Research Methods*, 49(1):335–362.
- James S. Kim, Mary A. Burkhauser, Jackie E. Relyea, Joshua B. Gilbert, Ethan Scherer, Jill Fitzgerald, Douglas Mosher, and Joseph McIntyre. 2023. [A longitudinal randomized trial of a sustained content literacy intervention from first to second grade: Transfer effects on students’ reading comprehension](#). *Journal of Educational Psychology*, 115(1):73–98.
- James S. Kim, Jackie E. Relyea, Mary A. Burkhauser, Ethan Scherer, and Patrick Rich. 2021. [Improving elementary grade students’ science and social studies vocabulary knowledge depth, reading comprehension, and argumentative writing: A conceptual replication](#). *Educational Psychology Review*, 33(4):1935–1964.
- Daniël Lakens. 2017. [Equivalence tests: A practical primer for \$t\$ tests, correlations, and meta-analyses](#). *Social Psychological and Personality Science*, 8(4):355–362.
- Daniël Lakens, Anne M. Scheel, and Peder M. Isager. 2018. [Equivalence testing for psychological research: A tutorial](#). *Advances in Methods and Practices in Psychological Science*, 1(2):259–269.
- Sora Lee and Daniel M. Bolt. 2018. [An alternative to the 3PL: Using asymmetric item characteristic curves to address guessing effects](#). *Journal of Educational Measurement*, 55:90–111.
- Xiangyi Liao and Daniel M. Bolt. 2021. [Item characteristic curve asymmetry: A better way to accommodate slips and guesses than a four-parameter model?](#) *Journal of Educational and Behavioral Statistics*, 46(6):753–775.
- Yang Liu and Xiaojing Wang. 2020. [Bayesian nonparametric monotone regression of dynamic latent traits in item response theory models](#). *Journal of Educational and Behavioral Statistics*, 45(3):274–296.
- Eric Loken and Kelly L. Rulison. 2010. [Estimation of a four-parameter item response theory model](#). *British Journal of Mathematical and Statistical Psychology*, 63(3):509–525.
- George G. Lorentz. 1953. *Bernstein Polynomials*. University of Toronto Press, Toronto.
- Sean Patrick McCarron and Victor Kuperman. 2021. [Is the author recognition test a useful metric for native and non-native English speakers? An item response theory analysis](#). *Behavior Research Methods*, 53(5):2226–2237.
- Shakked Noy and Whitney Zhang. 2023. [Experimental evidence on the productivity effects of generative artificial intelligence](#). *Science*, 381(6654):187–192.
- Sonia Petrone. 1999. [Random Bernstein polynomials](#). *Scandinavian Journal of Statistics*, 26(3):373–393.
- Nicholas G. Polson, James G. Scott, and Jesse Windle. 2013. [Bayesian inference for logistic models using Pólya–Gamma latent variables](#). *Journal of the American Statistical Association*, 108(504):1339–1349.
- J. O. Ramsay. 1988. [Monotone regression splines in action](#). *Statistical Science*, 3(4):425–441.
- J. O. Ramsay. 1991. [Kernel smoothing approaches to nonparametric item characteristic curve estimation](#). *Psychometrika*, 56(4):611–630.
- James S. Roberts, John R. Donoghue, and James E. Laughlin. 2000. [A general item response theory model for unfolding unidimensional polytomous responses](#). *Applied Psychological Measurement*, 24(1):3–32.
- Fumiko Samejima. 2000. [Logistic positive exponent family of models: Virtue of asymmetric item characteristic curves](#). *Psychometrika*, 65:319–335.
- Michelle Vaccaro, Abdullah Almaatouq, and Thomas Malone. 2024. [When combinations of humans and AI are useful: A systematic review and meta-analysis](#). *Nature Human Behaviour*, 8:2293–2303.
- Mike Wu, Richard L. Davis, Benjamin W. Domingue, Chris Piech, and Noah Goodman. 2020. [Variational item response theory: Fast, accurate, and expressive](#). In *Proceedings of the 13th International Conference on Educational Data Mining (EDM)*. ArXiv:2002.00276.

A Experimental Details

Cross-validation. Folds were formed with a fixed seed by randomly assigning each student’s responses to ten folds, with a random per-student off-set so that the remainder of the student’s response count does not always land on the same fold. Every student therefore appears in every training fold and (unless they answered fewer than ten items) in every test fold, so the ability distribution is identical across folds by construction; Table 5 reports the balance of the resulting folds in response rate and size. Folds are shared across models, so all comparisons are paired by fold.

Dataset	Resp. / student	Held out / fold	%-correct spread	Size imbalance	In all folds
art	50	5	0.93 pp	0.00%	100%
spelling	101–109	10–11	1.96 pp	0.26%	100%
malawi_math	1–10	1	2.19 pp	0.15%	98.1%
online_rct	1–30	1–3	0.89 pp	0.14%	98.9%
content_lit	20	2	2.90 pp	0.00%	100%
vocab	12–24	1–3	1.47 pp	0.96%	100%

Table 5: Balance of the ten cross-validation folds used in Section 4.1. “Held out / fold” is the number of a student’s responses in each test fold; “%-correct spread” is the range of the held-out percentage correct across the ten folds (percentage points); “size imbalance” is the range of held-out fold sizes relative to their mean; “in all folds” is the share of students who appear in every test fold (the remainder answered fewer than ten items).

Dataset	Logistic (2PL)		Normal-ogive (2PNO)		I-spline		Free-form: bounded – monotone	
	$n=10$	$n=20$	$n=10$	$n=20$	$n=10$	$n=20$	$n=10$	$n=20$
art	+1.97	+1.38	+0.15	−0.14	+2.04	+0.14	−1.78	−8.66 [†]
spelling	+0.35	+0.02	−0.20	−0.55	+0.14	+0.06	−0.13	+0.10
malawi_math	+0.78	+0.46	+0.26	+0.25	+2.94	+1.25	−3.51	−4.57
online_rct	+0.49	+0.19	+0.14	−0.02	+3.24	+1.02	−1.63	−1.22
content_lit	+0.53	+0.15	+0.43	+0.07	+0.18	−0.06	−0.25	−1.03
vocab	+1.14	+0.86	+0.51	+0.48	+0.18	−0.27	−2.21	−0.85

Table 6: Per-dataset differences in held-out NLL ($\times 10^{-3}$ nats; mean over the ten folds). The first six columns are bridged (exact closed form) minus direct for each IRF family and degree; every cell is TOST-equivalent at the 0.01-nat margin after BH correction, the largest gap being 0.003 nats. The last two columns are the flexible free-form model minus the monotone free-form model, both in quadrature form: relaxing monotonicity never hurts by more than 0.0001 nats and never helps by more than 0.009 nats. [†]The only cell not TOST-equivalent ($q = 0.14$).

Training (RQ-1). All models were trained with Adam (learning rate 0.02) on mini-batches of 256 students for at most 100 epochs, with early stopping on the log-loss of a validation slice (12.5% of each student’s training responses, drawn at random within student; patience 10, minimum improvement 10^{-4}) and the best checkpoint restored. Direct models used 80 Gauss–Hermite nodes under a $\mathcal{N}(0, 1)$ prior; bridged models used the exact fold under the Beta(4, 4) prior with $h = 3$ (Section 5), and the quadrature-form results in Table 6 used 80 Gauss–Jacobi nodes. The 2PL and 2PNO families have a discrimination and a difficulty per item; the I-spline family generates the Bernstein weights from a monotone I-spline with 10 basis functions of order 2 per item; the free-form families parameterize the item-by- $(n+1)$ weight matrix as a rank-2 factorization (a two-dimensional item embedding times a shared $2 \times (n+1)$ basis) plus one shared $(n+1)$ -vector, and enforce monotonicity, where applicable, by chaining the weights through sigmoid increments. The parameter counts in Table 3 are means over the six datasets.

Simulation (RQ-2). Each run draws 2,000 students with $\eta \sim \mathcal{N}(0, 1)$ clipped to $[-3, 3]$, so the 4PL’s Normal prior is correctly speci-

fied and the Bernstein models’ Beta(4, 4) prior is the mildly misspecified one. All items have $a = 1.6$; the standard items have $b \in \{-1.2, -0.4, 0.4, 1.2\}$ and the AI-mediated items $b \in \{-0.4, 0, 0.4, 0.8, 1.2, 1.6\}$. The Bernstein models use degree $n = 12$; the 4PL uses 48 Gauss–Hermite nodes. Every model is fit by full-batch Adam (learning rate 0.05, 200 epochs). RMSE is computed on a 200-point grid uniform in η , and AIC/BIC use the in-sample marginal likelihood with $N = 2,000$ response patterns. At a milder AI effect ($A_j = 0.92$, $s_r = 0.80$), the flexible model’s held-out NLL advantage over both monotone base-lines shrinks from roughly 0.08–0.10 to 0.045 nats per student but persists in all 20 runs.