

Bayesian Consensus Calibration of Continuously Evolving IRT Item Banks

Paul A. Jewsbury^{ID} and Steven W. Nydick^{ID} and Manqian Liao^{ID} and Siyuan (Marco) Chen^{ID}

Duolingo

{paul.jewsbury, steven, mancy, marco}@duolingo.com

Abstract

AI-based item generation and NLP-based prediction of item parameters are producing item banks that are substantially larger, sparser, and more frequently updated than conventional banks. Hierarchical Bayesian item response theory (IRT) is a natural calibration framework for such banks, but the common practice of refitting the entire accumulated response history at each update is costly and can exceed available memory. We describe *consensus calibration*, a divide-and-conquer procedure that calibrates each time period independently and reconstructs the pooled posterior in two layers. First, the posterior draws of each period are mapped to a common metric by a robust characteristic-curve linking (Haebara) that is solved separately for each draw, which propagates the uncertainty of the linking transformation into the linked posteriors. Second, the linked item posteriors are combined as a product of Gaussian densities from which the population prior contributed by each period is removed and a single prior—obtained by consensus across the per-period population posteriors—is reinstated. The correction targets the posterior dispersion, not only its location. As evidence for consensus calibration, we compare it to a pooled single-run analysis on a large operational assessment in terms of item-parameter recovery, an uncertainty-by-exposure diagnostic, and the ability distributions.

1 Introduction

Assessment programs increasingly generate items automatically and assign their parameters by prediction from item content rather than solely by pretesting. The Duolingo English Test, for instance, combines automatic generation, feature-based parameterization, and frequent recalibration of a continuously evolving bank (A. A. von Davier, 2024; Nydick et al., 2026a). The resulting response data

are large, sparse, and adaptively sampled. Reproducing what a single fit to all of this accumulated data would yield, but without refitting past data as the bank grows, is the problem this paper addresses.

Explanatory hierarchical Bayesian IRT models suit this setting: the prior on each item's parameters is an estimated population distribution conditional on item features (De Boeck and Wilson, 2004). Sparsely observed items are shrunk toward their feature-based predictions, and the posterior supplies the parameter uncertainty on which adaptive selection and scoring depend (Fox, 2010; Sharpnack et al., 2026). The operational difficulty lies in recalibration: standard practice refits a single model to all accumulated data at each update, so cost grows continuously and, at contemporary bank sizes, holding the full history in memory may exceed available computational resources (Nydick et al., 2026b).

We treat recalibration as a divide-and-conquer Bayesian computation (Xu et al., 2024). Each period is calibrated independently, and the period posteriors are combined without revisiting earlier data. Two features distinguish our case from the standard divide-and-conquer approach. First, each period's fit identifies the latent metric only up to a linear transformation, so the period posteriors are not directly comparable and must be linked. Second, the prior is not fixed but is itself estimated and hierarchical: each period's posterior incorporates its own population prior, which the combination must replace with a single prior to avoid understating uncertainty.

This paper makes four contributions: (i) a per-draw robust linking step that maps each period to a common metric, adopting the sampling strategy of Baldwin (2011) with a response-count-weighted robust characteristic-curve criterion across multiple periods, so that the uncertainty of the linking transformation is propagated into the period posteriors rather than treated as a fixed transformation with an

added error term; (ii) a two-layer consensus aggregation that combines the linked item posteriors and replaces the prior contributed by each period with a single prior obtained by consensus across the per-period population posteriors; (iii) an explicit identifiability condition with a population-prediction fallback for items not identified by the period data beyond the population model; and (iv) an empirical evaluation that compares consensus calibration with a pooled benchmark, illustrated on a large operational assessment.

2 Background and Related Work

Automatic item generation increasingly relies on large language models, resulting in item banks much larger than was possible with manual authoring (Gierl and Haladyna, 2013; M. von Davier, 2018; A. A. von Davier, 2024). Independently, NLP models predict item difficulty and related parameters from item content—using engineered text features, pretrained-language-model embeddings, or automated machine learning over both—so that items can be operationalized with little or no pretesting (Settles et al., 2020; McCarthy et al., 2021; Yancey et al., 2024; Sharpnack et al., 2024). Benedetto et al. (2023) and AlKhuyaey et al. (2024) survey text-based difficulty estimation, which centers on language assessment but reaches content-knowledge domains such as computer science and medicine. A recent shared task, for instance, benchmarked such systems on high-stakes medical items (Yaneva et al., 2024). These methods are deployed operationally, where feature-based parameters let newly generated items enter scoring before substantial response data accrue (A. A. von Davier, 2024; Nydick et al., 2026a,b).

In hierarchical Bayesian IRT, item parameters are drawn from a population distribution whose hyperparameters are estimated jointly with the items (Fox, 2010). Explanatory IRT generalizes the population model to a regression of item parameters on item features (De Boeck and Wilson, 2004). This regression is the object that feature-based parameter prediction estimates and that Bayesian calibration engines implement at scale (e.g., Nydick et al., 2026b), so the prior in such a calibration is the feature-conditional distribution of the item parameters.

Placing separately calibrated forms on a common metric is the problem of linking and equating, commonly solved by characteristic-curve methods

that estimate a linear transformation by matching model-implied curves on common items (Haebara, 1980; Stocking and Lord, 1983; Kolen and Brennan, 2014). Exact invariance of the linking items across calibrations is neither realistic nor a prerequisite for valid comparison (Robitzsch and Lüdtke, 2023). Characteristic-curve criteria fit the transformation to the anchors in aggregate, and robust variants attenuate the influence of the minority of items that function differently (Robitzsch, 2024), suiting them to this partial-invariance context. Because the link is itself estimated from the same data as the item parameters, linking error is generally dependent on sampling and measurement error. Jewsbury (2025a) proposes and validates a taxonomy of variance-estimation methods that propagate linking error while accounting for these dependencies (Jewsbury, 2025b, 2026). Baldwin (2011) carries this error in the posterior sample itself: the transformation is estimated from a single draw of each calibration and applied to that draw, repeated until the transformed posterior is assembled. We adopt this strategy for the multi-period case. This complements Bayesian nonparametric equating, which links observed score distributions (Karabatsos and Walker, 2009), by performing IRT-parameter linking in a Bayesian setting.

Combining posteriors computed on disjoint subsets of a single dataset is the subject of consensus Monte Carlo (Scott et al., 2016) and embarrassingly parallel MCMC (Neiswanger et al., 2014), which parallelize one fit by partitioning the data across workers and merging the subset posteriors as a product of densities (Hinton, 2002; Tresp, 2000). For IRT, Xu et al. (2024) apply this within a single calibration, randomly partitioning the examinee sample and merging the subset posteriors by a Wasserstein barycenter, with the identification constraints held fixed so that no linking is required.

Periodic recalibration is a different problem. Applied to an evolving bank, these single-fit methods would re-analyze all the data at each update, treating it as one dataset to be partitioned and fit together rather than a sequence of runs to be combined. The proposed method instead calibrates each period once and combines a new period with the already-computed posteriors of the earlier runs, so an update’s cost scales with the new data rather than the full history. This difference forces two departures from the single-calibration setting: (a) because the periods are fit separately rather than within one run, they do not share a latent metric and

must be linked; and (b) because each period is its own hierarchical fit, it carries an estimated, period-specific prior that the combination must reconcile to a single prior. Consensus calibration addresses both.

3 Setting and Notation

Let $m = 1, \dots, M$ index time periods, each calibrated independently by a Bayesian explanatory IRT engine (Nydicke et al., 2026b). The calibration of period m returns posterior draws of each item’s parameter vector ξ_j and of the population hyperparameters of each item group: the coefficients β of a multivariate regression of item parameters on item features and a residual covariance Σ , so that the population model is $\xi_j \sim \mathcal{N}(\mathbf{X}_j\beta, \Sigma)$ given features $\mathbf{X}_j$. Each item is assumed to load on a single latent dimension, which fixes the orientation of the latent space; item responses then determine each dimension only up to its origin and unit, so the period- m metric is identified only up to a linear transformation,

$$\theta_{\text{ref}} = A_m \theta + B_m, \quad A_m > 0, \quad (1)$$

with constants specific to the period and dimension (the dimension index is suppressed). A designated reference period has $(A_m, B_m) = (1, 0)$. The objective is to reconstruct the posterior that a single analysis pooling all periods (i.e., the full dataset) would yield.

The method uses only two properties of the item parameterization. First, each item j has an expected-score (characteristic) curve $E_j(\theta)$, the expected item score at latent value θ ; matching these curves on anchor items is what estimates the linking constants (Section 4.1). Second, the change of metric (1) induces a transformation of ξ_j ; applying the transformation draw-wise is what carries every item’s posterior onto the reference metric.

Assumptions. The procedure relies on four assumptions. (i) *Disjoint examinee samples*, so that the period likelihoods are conditionally independent and the pooled likelihood factors across periods (Section 4.2). (ii) *Simple structure*: each item loads on a single latent dimension, reducing the metric indeterminacy to the linear map (1). (iii) *Aggregate anchor invariance*: the metric change is captured by (A_m, B_m) and the anchor set is invariant in aggregate (Section 4.1). (iv) *Gaussian summaries*: each period’s item posterior, prior con-

tribution, and population-coefficient posterior are summarized by Gaussian moments (Appendix A).

4 Method

We aggregate independently calibrated periods rather than updating sequentially—carrying each period’s posterior forward as the next period’s prior—for three reasons. First, fitting each period as a self-contained hierarchical model lets its own population prior stabilize its sparsely observed items, rather than a prior whose form was fixed in advance. Second, the model specification, including the item parameter prediction model, may change between periods. Third, aligning the periods needs weaker invariance than sequential updating: its reused item priors presume each item is invariant, whereas a transformation fit to the anchors collectively needs only the anchor set to hold in aggregate (Robitzsch and Lüdtke, 2023).

4.1 Per-draw robust linking

Under (1), the expected-score curve of item j expressed on the reference metric is

$$E_j^{m \rightarrow \text{ref}}(\theta) = E_j((\theta - B_m)/A_m), \quad (2)$$

so the constants (A_m, B_m) determine a transformation of the item parameters (for example, locations map as $d \mapsto A_m d + B_m$ and log-discriminations as $\log a \mapsto \log a - \log A_m$). The linking constants for period m are chosen to minimize the weighted discrepancy between the reference curves and the transformed period- m curves over an anchor set $\mathcal{A}$. Writing the discrepancy at θ as $D_j(\theta) = E_j^{\text{ref}}(\theta) - E_j^{m \rightarrow \text{ref}}(\theta)$,

$$H_m(A, B) = \int \sum_{j \in \mathcal{A}} \omega_j \rho(D_j(\theta)) w(\theta) d\theta, \quad (3)$$

where w is a weight on the latent metric and ω_j is proportional to the number of responses on which anchor j was calibrated. Weighting anchors by their response count lets better-estimated items contribute more to the link, approximating concurrent (joint) calibration, in which items with more data exert more influence on the common metric. The integral is evaluated by quadrature. The loss ρ is a pseudo-Huber loss function, a smooth surrogate for the least-absolute (L_1) criterion of robust Haebara linking (Robitzsch, 2024). Note that any of the moment or characteristic-curve linking methods unified by M. von Davier and A. A. von Davier (2007) could replace Haebara here.

The criterion (3) is minimized separately for each posterior draw, in the manner of Baldwin (2011), rather than once at the posterior mean. For draw s , the anchor curves are evaluated at that draw’s parameter values and (3) is solved to obtain $(A_m^{(s)}, B_m^{(s)})$; that draw’s transformation is then applied to draw s of every item, anchors and non-anchors alike. The collection $\{(A_m^{(s)}, B_m^{(s)})\}_{s=1}^S$ is a sample from the posterior of the linking constants. Applying it draw-wise propagates the linking uncertainty, together with its dependence on the item parameters, into the linked posteriors. Within the posterior, draw-wise linking plays the role that dependence-aware variance estimation plays for frequentist linking, where treating linking error as independent of the other sources of error misstates the uncertainty of downstream comparisons (Jewsbury, 2025a, 2026).

Two filters are applied to the anchor set $\mathcal{A}$ before linking. The first removes any anchor whose potential scale reduction factor $\hat{R}$ (Gelman and Rubin, 1992), maximized over the item’s parameters, exceeds 1.05 in either the period being linked or the reference period. The second is an iterative purification (Candell and Drasgow, 1988): within each dimension, a provisional mean–sigma link (Kolen and Brennan, 2014) is fit to the anchors’ trait-metric locations (posterior means), anchors whose absolute residuals from the provisional link fall in the top decile are removed, and the link is refit, for at most five iterations. Here the reference period denotes the metric-defining calibration, not the pooled benchmark of Section 5.

4.2 Consensus aggregation and prior de-duplication

Linking places each item’s period posteriors on the reference metric; aggregation now combines them into the pooled posterior, first as a precision-weighted product and then with a correction for the prior that each period contributes. Consider one item, with its index suppressed, and let its linked posterior in period m be summarized by a Gaussian with mean $\hat{\mu}_m$ and precision $\Lambda_m = \hat{\Sigma}_m^{-1}$. Sums over m run over the periods in which the item was observed. If the periods comprise disjoint examinee samples, the joint likelihood factors across periods given the item parameters, and the pooled posterior is proportional to the product of the period posteriors with the multiply-counted prior removed. For the Gaussian summaries the product is

Gaussian, with

$$\tilde{\Lambda} = \sum_m \Lambda_m, \quad \tilde{\mu} = \tilde{\Lambda}^{-1} \sum_m \Lambda_m \hat{\mu}_m. \quad (4)$$

Each factor in (4) incorporates the population prior, so the product incorporates it M times and is over-concentrated. Let the prior contributed by period m be Gaussian with precision Ω_m and mean $\mathbf{m}_m$ —the moments of that period’s population prediction for the item, mapped to the reference metric by the same draw-wise transformation—and let the single replacement prior have precision Ω^* and mean $\mathbf{m}^*$. The corrected natural parameters are

$$\begin{aligned} \Lambda_c &= \sum_m \Lambda_m - \sum_m \Omega_m + \Omega^*, \\ \mathbf{b}_c &= \sum_m \Lambda_m \hat{\mu}_m - \sum_m \Omega_m \mathbf{m}_m + \Omega^* \mathbf{m}^*, \end{aligned} \quad (5)$$

with consensus mean $\mu_c = \Lambda_c^{-1} \mathbf{b}_c$ (Appendix A). When the periods share a fixed prior, (5) reduces to $\Lambda_c = \sum_m \Lambda_m - (M-1)\Omega$, the combination rule of the Bayesian committee machine (Tresp, 2000). The general form of (5) is required because the prior is estimated and differs across periods.

The replacement prior is obtained by consensus at the population layer. Let the period- m posterior of the population coefficients β , mapped to the reference metric, be approximately Gaussian with mean $\mathbf{b}_m$ and covariance $\mathbf{V}_m$. Under a weak zero-mean Gaussian hyperprior with precision τ , the consensus coefficients are the Gaussian product of the per-period posteriors with the hyperprior counted once,

$$\beta^* = \left(\sum_m \mathbf{V}_m^{-1} - (M-1)\tau\mathbf{I} \right)^{-1} \sum_m \mathbf{V}_m^{-1} \mathbf{b}_m. \quad (6)$$

The consensus residual covariance Σ^* is the item-count-weighted mean of the per-period residual covariances taken in log-Cholesky coordinates—the Fréchet mean under the log-Cholesky metric (Lin, 2019; Pinheiro and Bates, 1996)—which is positive definite by construction (Appendix B). The replacement prior of (5) for the item under consideration is then its consensus population prediction: $\mathbf{m}^* = \mathbf{X}\beta^*$, with $\mathbf{X}$ its features, and $\Omega^* = (\Sigma^*)^{-1}$. The population layer presupposes a common population-model specification across periods, as in our application; under specification changes, the replacement prior can instead be formed from the item-level predictions. Consensus thus applies at both layers—the item layer, (4)–(5),

and the population layer, (6)—and the prior correction restores the dispersion that the naive product understates. The aggregate is formed analytically: (5) is evaluated in closed form and the resulting Gaussian is sampled to provide the item’s consensus draws.

4.3 Eligibility and the population fallback

The subtraction in (5) can render Λ_c non-positive-definite, as tends to happen for items that the period data barely identify beyond the population model or whose Gaussian summaries are noisy. Rather than regularize Λ_c (for example by pseudo-inversion), which would impute dispersion the data do not support, we flag these items and assign them the population prediction $\xi_j \sim \mathcal{N}(\mathbf{X}_j\beta^*, \Sigma^*)$ —the replacement prior of (5). The proportion of items routed to the fallback is a diagnostic of whether the periods are individually informative enough for the procedure.

5 Empirical Evaluation

5.1 Data and configuration

The procedure was applied to operational response data from the Duolingo English Test, a high-volume, continuously evolving language assessment (Nydict et al., 2026a; Naismith et al., 2025), measuring two latent dimensions with items in several response formats (selected-response, partial-credit, and continuous). Each period was calibrated by a scalable Bayesian explanatory IRT engine (Nydict et al., 2026b) with period-specific item posteriors thinned to approximately 1,000 draws. The run comprised four consecutive quarterly calibration periods spanning approximately one year with the third period defining the reference metric. Each period included on the order of tens of thousands of items and hundreds of thousands of examinees. Linking used the robust per-draw criterion of Section 4.1 with the $\hat{R}$ -and-purification anchor screen, solved for 200 draws of each non-reference period; the transformed draws form that period’s linked sample. Aggregation used the analytic consensus method of Section 4.2 with the prior correction: each period’s Gaussian summary was estimated from its linked draws, and the consensus Gaussian was sampled to provide 2,000 draws per item.

5.2 Validation protocol

To obtain evidence that the aggregate reproduces a single pooled analysis, we fit one here as a bench-

mark. To avoid confounding the results, the benchmark is fit with a separate ability distribution per period—a multiple-group model in which each period’s examinees form a group—matching the ability structure of the per-period calibrations.

All comparisons are restricted to converged item parameters: an (item, parameter) pair enters only if its split-chain $\hat{R}$ does not exceed 1.1 in the pooled run and in every period run containing the item, since posterior summaries that have not converged on at least one side of the comparison are not interpretable. This screen removed 1.9% of (item, parameter) pairs.

Three comparisons are reported. (i) *Recovery of item parameters*: the correlation and root-mean-square error (RMSE) of the posterior means, and the same statistics for the posterior standard deviations (Figure 1; Table 1). Agreement of the means indicates unbiased linking and aggregation; agreement of the standard deviations indicates that the posterior dispersion, and not only its location, is recovered. (ii) *Uncertainty by exposure*: because prior over-counting most affects items observed sparsely within each period, the ratio of posterior standard deviations, $SD_{\text{cons}}/SD_{\text{pool}}$, is examined within tertiles of per-period response volume (Table 1, SD-ratio rows); the signature of over-counting is a ratio below one at low exposure that approaches one as exposure increases, whereas a ratio above one indicates residual linking variability. (iii) *Ability-distribution agreement*: because the item link and the ability-metric transformation share the same constants, comparing each period’s ability mean and standard deviation checks the linking independently of the items.

5.3 Results

Figure 1 plots the consensus against the pooled posterior, item by item, for the posterior mean (top) and the posterior standard deviation (bottom), with points colored by per-period exposure tertile; Table 1 summarizes the agreement. Posterior means agreed closely ($r = .998$ for d and $.991$ for $\log a_1$; RMSE 0.129 and 0.081), indicating that the per-draw linking and the precision-weighted aggregation introduce no systematic distortion of location. Posterior standard deviations also agreed ($r = .970$ and $.920$; RMSE 0.066 and 0.036), evidence that the prior correction recovers most of the dispersion of the pooled posterior and not only its location; the SD-ratio rows quantify the remaining shortfall.

The exposure diagnostic of comparison (ii)

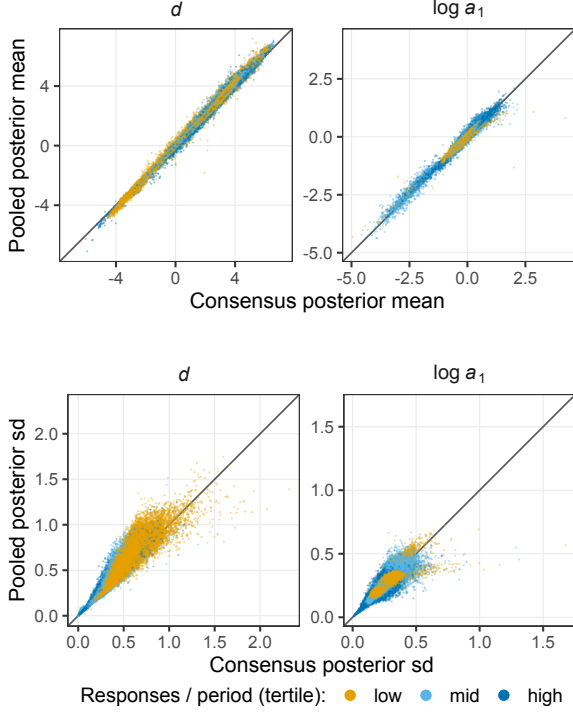


Figure 1: Item-level agreement between the consensus aggregate and the pooled benchmark, for item parameters that converged in the pooled run and in every period run ($\hat{R} \leq 1.1$). *Top*: posterior means; *bottom*: posterior standard deviations; columns are item parameters; the gray line is the identity. Color gives per-period exposure tertile.

appears in the SD-ratio rows of Table 1: 0.91/0.94/0.94 for d and 0.96/0.95/0.98 for $\log a_1$ across low, mid, and high per-period exposure. Both parameters show mild under-dispersion in the direction expected of residual prior over-counting, particularly in the lowest-exposure tertile.

For comparison (iii), the per-period ability means implied by the linked period calibrations differed by at most 0.08 reference-metric units, and the ratio of the corresponding ability standard deviations ranged from 0.98 to 1.07.

6 Discussion

Consensus calibration is an instance of computational psychometrics, bringing scalable Bayesian computation to hierarchical IRT calibration (A. A. von Davier et al., 2022). It reconstructs the pooled posterior from independent per-period fits. The periods are linked by a robust characteristic-curve criterion solved separately for each posterior draw, in the manner of Baldwin (2011), which propagates

		d	$\log a_1$
Post. mean	r	0.998	0.991
	RMSE	0.129	0.081
Post. SD	r	0.970	0.920
	RMSE	0.066	0.036
SD ratio	low expo.	0.91	0.96
	mid expo.	0.94	0.95
	high expo.	0.94	0.98

Table 1: Agreement of the consensus aggregate with the pooled benchmark, for converged item parameters only: correlation r and RMSE between the consensus and pooled posterior means and posterior standard deviations, and the mean posterior-SD ratio (consensus/pooled) within low/mid/high tertiles of per-period exposure. A ratio below one indicates residual under-dispersion from prior over-counting.

the uncertainty of the linking transformation. The linked posteriors are then combined, with the prior contributed by each period removed and a single prior obtained by consensus reinstated. The cost of an update consequently scales with the new period rather than the accumulated history, and the full history need not be held in memory at once. In the benchmark comparison, the consensus posterior means and standard deviations agreed closely with those of the pooled fit.

The procedure builds on feature-based parameter prediction rather than replacing it: the population model aggregated at its second layer is the regression of item parameters on item features that such prediction estimates, so calibration refines those predictions rather than discarding them. Because the link is solved per draw, the linked posteriors carry the linking error, together with its dependence on the item parameters, forward as calibrated uncertainty (Nydicke et al., 2026a; Jewsbury, 2025a).

The procedure has limitations. Cross-period conditional independence requires disjoint examinee samples, since substantial overlap would count some responses more than once. The factors combined are plug-in Gaussian summaries of each period’s hierarchical posterior, an approximation whose validity was assessed empirically in the benchmark comparison. Finally, the aggregation targets each item’s marginal posterior; the joint dependence across items induced by shared hyperparameters and common linking constants is not reconstructed.

References

- Samah AlKhazaey, Floriana Grasso, Terry R. Payne, and Valentina Tamma. 2024. [Text-based question difficulty prediction: A systematic review of automatic approaches](#). *International Journal of Artificial Intelligence in Education*, 34:862–914.
- Peter Baldwin. 2011. [A strategy for developing a common metric in Item Response Theory when parameter posterior distributions are known](#). *Journal of Educational Measurement*, 48(1):1–11.
- Luca Benedetto, Paolo Cremonesi, Andrew Caines, Paula Buttery, Andrea Cappelli, Andrea Giussani, and Roberto Turrin. 2023. [A survey on recent approaches to question difficulty estimation from text](#). *ACM Computing Surveys*, 55(9):178:1–178:37.
- Gregory L. Candell and Fritz Drasgow. 1988. [An iterative procedure for linking metrics and assessing item bias in item response theory](#). *Applied Psychological Measurement*, 12(3):253–260.
- Paul De Boeck and Mark Wilson, editors. 2004. *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*. Springer, New York.
- Jean-Paul Fox. 2010. *Bayesian Item Response Modeling: Theory and Applications*. Springer, New York.
- Andrew Gelman and Donald B. Rubin. 1992. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472.
- Mark J. Gierl and Thomas M. Haladyna, editors. 2013. *Automatic Item Generation: Theory and Practice*. Routledge, New York.
- Tomokazu Haebara. 1980. Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research*, 22(3):144–149.
- Geoffrey E. Hinton. 2002. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800.
- Paul A. Jewsbury. 2025a. [Generally applicable variance estimation methods for common-population linking](#). *Journal of Educational and Behavioral Statistics*, 50(4):651–681.
- Paul A. Jewsbury. 2025b. [Standard error estimation for subpopulation non-invariance](#). *Applied Psychological Measurement*, 49(8):477–505.
- Paul A. Jewsbury. 2026. [Linking error on achievement levels accounting for dependencies and complex sampling](#). *Journal of Educational Measurement*, 63(1):e12439.
- George Karabatsos and Stephen G. Walker. 2009. [A Bayesian nonparametric approach to test equating](#). *Psychometrika*, 74(2):211–232.
- Michael J. Kolen and Robert L. Brennan. 2014. *Test Equating, Scaling, and Linking: Methods and Practices*, 3rd edition. Springer, New York.
- Zhenhua Lin. 2019. [Riemannian geometry of symmetric positive definite matrices via Cholesky decomposition](#). *SIAM Journal on Matrix Analysis and Applications*, 40(4):1353–1370.
- Arya D. McCarthy, Kevin P. Yancey, Geoffrey T. LaFlair, Jesse Egbert, Manqian Liao, and Burr Settles. 2021. [Jump-starting item parameters for adaptive language tests](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 883–899, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Ben Naismith, Ramsey Cardwell, Geoffrey T. LaFlair, Steven W. Nydick, and Masha Kostromitina. 2025. [Duolingo English Test: Technical manual](#). Duolingo research report, Duolingo.
- Willie Neiswanger, Chong Wang, and Eric P. Xing. 2014. Asymptotically exact, embarrassingly parallel MCMC. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 623–632, Arlington, VA. AUAI Press.
- Steven W. Nydick, Manqian Liao, Siyuan Chen, and J. R. Lockwood. 2026a. Maintaining the meaning of test scores in a high-volume evolving assessment. In Ben Naismith, Alina A. von Davier, Jill Burstein, and Geoffrey T. LaFlair, editors, *The Routledge Handbook of Digital Language Assessment: Innovations and Insights from the Duolingo English Test*. Routledge. In press.
- Steven W. Nydick, Manqian Liao, and J. R. Lockwood. 2026b. [A scalable parametric item calibration engine \(SPICE\) for explanatory IRT with sparse data](#). arXiv:2605.21782 [stat.ME].
- José C. Pinheiro and Douglas M. Bates. 1996. Unconstrained parametrizations for variance-covariance matrices. *Statistics and Computing*, 6(3):289–296.
- Alexander Robitzsch. 2024. [sirt: Supplementary Item Response Theory Models](#). R package; implements robust Haebara linking with L_p loss.
- Alexander Robitzsch and Oliver Lüdtke. 2023. [Why full, partial, or approximate measurement invariance are not a prerequisite for meaningful and valid group comparisons](#). *Structural Equation Modeling: A Multidisciplinary Journal*, 30(6):859–870.
- Steven L. Scott, Alexander W. Blocker, Fernando V. Bonassi, Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. 2016. [Bayes and big data: The consensus Monte Carlo algorithm](#). *International Journal of Management Science and Engineering Management*, 11(2):78–88.
- Burr Settles, Geoffrey T. LaFlair, and Masato Hagiwara. 2020. [Machine learning-driven language assessment](#). *Transactions of the Association for Computational Linguistics*, 8:247–263.

James Sharpnack, Phoebe Mulcaire, Klinton Bicknell, Geoffrey T. LaFlair, and Kevin P. Yancey. 2024. [AutoIRT: Calibrating item response theory models with automated machine learning](#). arXiv:2409.08823 [cs.LG].

James Sharpnack, Steven W. Nydick, J. R. Lockwood, Alexander Tsigler, and Alina A. von Davier. 2026. Calibration, scoring, and administration of a high-stakes computerized adaptive test. In Ben Naismith, Alina A. von Davier, Jill Burstein, and Geoffrey T. LaFlair, editors, *The Routledge Handbook of Digital Language Assessment: Innovations and Insights from the Duolingo English Test*. Routledge. In press.

Martha L. Stocking and Frederic M. Lord. 1983. Developing a common metric in item response theory. *Applied Psychological Measurement*, 7(2):201–210.

Volker Tresp. 2000. A Bayesian committee machine. *Neural Computation*, 12(11):2719–2741.

Alina A. von Davier, Robert J. Mislevy, and Jiangang Hao, editors. 2022. *Computational Psychometrics: New Methodologies for a New Generation of Digital Learning and Assessment, with Examples in R and Python*. Methodology of Educational Measurement and Assessment. Springer, Cham, Switzerland.

Alina A. von Davier, Andrew Runge, Yena Park, Yigal Attali, Jacqueline Church, and Geoffrey T. LaFlair. 2024. The item factory: Intelligent automation in support of test development at scale. In Hong Jiao and Robert W. Lissitz, editors, *Machine Learning, Natural Language Processing, and Psychometrics*, pages 1–25. Information Age Publishing, Charlotte, NC.

Matthias von Davier. 2018. [Automated item generation with recurrent neural networks](#). *Psychometrika*, 83(4):847–857.

Matthias von Davier and Alina A. von Davier. 2007. [A unified approach to IRT scale linking and scale transformations](#). *Methodology*, 3(3):115–124.

Sainan Xu, Jing Lu, Jiwei Zhang, Chun Wang, and Gongjun Xu. 2024. [Optimizing large-scale educational assessment with a “divide-and-conquer” strategy: Fast and efficient distributed Bayesian inference in IRT models](#). *Psychometrika*, 89(4):1119–1147.

Kevin P. Yancey, Andrew Runge, Geoffrey T. LaFlair, and Phoebe Mulcaire. 2024. BERT-IRT: Accelerating item piloting with BERT embeddings and explainable IRT models. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 428–438, Mexico City, Mexico. Association for Computational Linguistics.

Victoria Yaneva, Kai North, Peter Baldwin, Le An Ha, Saed Rezayi, Yiyun Zhou, Sagnik Ray Choudhury, Polina Harik, and Brian Clauser. 2024. Findings from the first shared task on automated prediction of difficulty and response time for multiple-choice

questions. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 470–482, Mexico City, Mexico. Association for Computational Linguistics.

A Derivation of the prior correction

All densities are over the item parameter vector ξ on the reference metric, and each period enters through the Gaussian summaries of Section 4.2: its linked item posterior p_m and prior contribution π_m are the Gaussian summaries, for which the algebra below is exact; relative to the hierarchical posterior, in which the prior is estimated and the periods share hyperparameters, the result is a plug-in approximation. The target posterior combines the period likelihoods with a single prior π^* , whereas the product of period posteriors carries the product of M priors; dividing out the per-period priors and multiplying in π^* gives

$$p^*(\xi) \propto \left[\prod_m p_m(\xi) \right] \frac{\pi^*(\xi)}{\prod_m \pi_m(\xi)}.$$

A Gaussian with mean $\mathbf{a}$ and precision $\mathbf{P}$ contributes $-\frac{1}{2}\xi^\top \mathbf{P} \xi + \xi^\top \mathbf{P} \mathbf{a}$ to the log density up to a constant. With period posteriors of precision Λ_m and mean $\hat{\mu}_m$, per-period priors of precision Ω_m and mean $\mathbf{m}_m$, and replacement prior $(\Omega^*, \mathbf{m}^*)$, collecting quadratic and linear terms yields $\Lambda_c = \sum_m \Lambda_m - \sum_m \Omega_m + \Omega^*$ and $\mathbf{b}_c = \sum_m \Lambda_m \hat{\mu}_m - \sum_m \Omega_m \mathbf{m}_m + \Omega^* \mathbf{m}^*$, with $\mu_c = \Lambda_c^{-1} \mathbf{b}_c$, valid whenever Λ_c is positive definite. Each quotient p_m/π_m is an unnormalized Gaussian factor whose precision $\Lambda_m - \Omega_m$ need not itself be positive definite; only Λ_c must be. Positive-definiteness is the eligibility condition of Section 4.3; the population fallback is used when it fails. Equal fixed priors recover $\Lambda_c = \sum_m \Lambda_m - (M-1)\Omega$, the combination rule of the Bayesian committee machine (Tresp, 2000); the same multiplicity of the prior is handled in consensus Monte Carlo and embarrassingly parallel MCMC by assigning each subset the prior raised to the power $1/M$ (Scott et al., 2016; Neiswanger et al., 2014). The correction above generalizes these devices to priors that are estimated, hierarchical, and different across periods, and to a replacement prior obtained by consensus.

B Log-Cholesky mean of covariances

Let a positive-definite Σ have lower Cholesky factor $\mathbf{L}$ with positive diagonal. Its log-Cholesky coordinates are the logarithm of the diagonal together with the strict lower-triangular entries of

$\mathbf{L}$, a smooth bijection onto an unconstrained Euclidean space whose inverse exponentiates the diagonal and returns $\mathbf{L}\mathbf{L}^\top$, necessarily positive definite (Pinheiro and Bates, 1996). A weighted arithmetic mean taken in these coordinates, with nonnegative weights summing to one, is the weighted Fréchet mean under the log-Cholesky metric on covariance matrices (Lin, 2019): it matches the unconstrained parameterization in which the covariances are estimated, is positive definite by construction, and has determinant equal to the weighted geometric mean of the period determinants, averaging dispersion on a multiplicative scale and avoiding the determinant inflation of the entrywise mean. Period m is weighted by its normalized item count in the block, and the coordinate ordering is held fixed across periods.