

# Exploratory Bayesian Multidimensional IRT with Rater Effects for Scoring Generative AI Usage

Leon Camus<sup>1</sup> , Sebastian Gombert<sup>1</sup>, Fabíola Ribeiro<sup>1</sup>, Gianluca Romano<sup>1</sup>,  
Carmen Köhler<sup>1</sup>, Hendrik Drachsler<sup>1,2</sup>

<sup>1</sup>DIPF | Leibniz-Institute for Research and Information in Education,

<sup>2</sup>Computer Science Department, Goethe University,

Correspondence: [l.camus@dipf.de](mailto:l.camus@dipf.de)

## Abstract

Generative AI is now embedded in how students produce academic work, yet the psychometric models used to score that work treat AI use as a single proficiency, collapsing it onto one ordinal scale. This conflates distinct interaction strategies: a student who treats a chatbot as a “research assistant” and one who delegates the thinking to it occupy different regions of a multidimensional space, not near and far points on one axis. We develop a multidimensional Bayesian graded response model that surfaces these differences without a pre-specified item-to-dimension mapping, with a doubly-centered rater-by-item submodel that separates rater style from item signal and a likelihood that accommodates mixed binary and polytomous rubric items. In-model identification constraints distort the posterior geometry for Hamiltonian Monte Carlo; we document three that fail in this setting: lower-triangular anchoring, an ordered positive diagonal, and a factorized loading matrix with automatic relevance determination (ARD). We instead sample the loading matrix unconstrained and identify it post-hoc via Varimax rotation and signed column-permutation matching (MatchAlign), a route not previously brought to multidimensional IRT with rater effects. Applied to roughly 18,000 ratings of 251 German undergraduates completing Critical Online Reasoning tasks, the model recovers three dimensions: active prompting effort, AI delegation of substantive content, and AI-delivered citation use. Working within one’s own academic domain shifts students toward more prompting and citation-seeking and away from delegation, consistent with a “research assistant” pattern of in-domain AI use.

## 1 Introduction

The widespread adoption of generative AI in education (Chan, 2023; Kasneci et al., 2023) has outpaced the psychometric frameworks used to evaluate student work produced with it (Kaldaras et al.,

2024). Standard rubric-based scoring typically assesses the quality of a submitted product without accounting for the conditions under which it was produced; Dawson et al. (2024) argue that this is a validity problem, since judgments about student capability are undermined when, because of AI usage, it is not clear anymore which parts of a given response were written by a student, and are therefore representative of assessed writing capabilities. It is therefore important to not merely assess test-takers’ products but also examine their interactions with AI models when assessing their writing capabilities. Existing frameworks address this by classifying the permitted or observed level of AI use on a single ordinal scale (Perkins et al., 2024). However, a single scale cannot usefully distinguish detailed interaction strategies that test-takers might adopt when using AI.

Distinguishing these strategies requires representing AI use as a profile across several behaviors rather than as a position on one scale. Consider two students who both rely heavily on a chatbot: one issues a sequence of self-formulated questions, asks for revisions, and writes their own conclusions while using the model mainly to locate sources; the other pastes the task prompt and adopts the output wholesale. On a single “amount of AI use” axis the two score almost identically, yet they differ sharply in what the AI actually contributed. A single ordinal scale can only order students from less to more AI use and cannot represent this difference in kind (Zhang et al., 2025; Dong et al., 2025); a multidimensional latent space instead places the two profiles in distinct regions. Recent log-based work (Güner and Er, 2025; McNichols et al., 2026; Ammari et al., 2026; Sawalha et al., 2024) and qualitative studies (Hadinejad et al., 2025; Kim et al., 2025) document that students adopt distinct interaction strategies rather than varying merely in proficiency: a student who treats a chatbot as a “research assistant” produces qualitatively different ar-

tifacts from one who treats it as a content provider, and whether one or the other is better cannot be solely captured by assessing the final product with a rubric. Aggregate scores and unidimensional IRT models absorb such profile differences into residual noise (Reckase, 2009).

In contrast to this, multidimensional item response theory allows for the expression of more detailed usage profiles (Reckase, 2009; Geweke and Zhou, 1996). A multidimensional latent space lets the data express qualitative differences in AI use without committing a priori to a single skill axis, and an explicit rater submodel helps separate rater severity and doubly-centered rater-by-item interactions (Cronbach et al., 1972; Linacre, 1989) from item-level signal. However, this setting requires accommodating mixed polytomous and binary items in a single rubric, which complicates likelihood estimation.

This model class is challenging for Hamiltonian Monte Carlo: the loading matrix is identified only up to rotation, sign, and column permutation, and in-model constraints that remove these symmetries distort the posterior geometry NUTS must traverse. We tried three standard constraints and found each either underidentified the model or induced local modes the sampler could not cross (Appendix B). We instead sample the loading matrix unconstrained and resolve all three symmetries post-hoc via Varimax rotation (Kaiser, 1958) and signed column-permutation matching (Poworoznek et al., 2025). To our knowledge no existing model combines full HMC sampling, post-hoc rotational identification, a doubly-centered rater submodel, and a mixed binary/polytomous IRT likelihood.

Our contributions are threefold. First, we introduce a **multidimensional graded response model** for ordinal rating data identified post-hoc via MatchAlign, with a doubly-centered full rater-by-item interaction and item-level local-scale loading shrinkage. Second, we provide a **negative-results documentation** of in-model identification approaches that fail in this setting, providing concrete guidance against parameterizations a practitioner might otherwise reach for first. Third, we use the model to carry out an **exemplary analysis of AI usage strategies of test-takers** during an online reasoning assessment. Our model can identify active prompting effort, AI delegation of substantive content, and AI-delivered citation use as distinctive factors between different usage profiles. It reports paired domain-match effects on these dimensions

consistent with a “research assistant” interpretation of in-domain AI use.

## 2 Related Work

The graded response model (Samejima, 1969) and its multidimensional extensions (Reckase, 2009) form the backbone of psychometric scoring for ordinal data, with Bayesian MCMC estimation of multidimensional IRT established by Béguin and Glas (2001). The classical approach to identifying the loading matrix is the lower-triangular constraint of Geweke and Zhou (1996); more recent work uses informative priors to induce soft sparsity (Ročková and George, 2016; Frühwirth-Schnatter et al., 2025). Global-local shrinkage priors on factor loadings, of which our  $\tau_A \lambda_{A,j}$  decomposition is an instance, originate with the multiplicative gamma process of Bhattacharya and Dunson (2011).

A separate line of research dispenses with in-model identification entirely and resolves rotation, sign, and permutation symmetries post-hoc: Stephens (2000) introduces relabelling algorithms for column- and label-symmetry in mixtures; Abmann et al. (2016) apply ex-post Procrustes matching to factor models; Papastamoulis and Ntzoufras (2022) introduce the RSP algorithm; and Poworoznek et al. (2025) introduce MatchAlign with Varimax rotation (Kaiser, 1958) and a greedy column-matching procedure. The combination of parameter-expanded rotation to sparsity with a mixed binary and ordinal IRT likelihood is developed independently by Li et al. (2025) using a Bayesian EM algorithm; our work uses full posterior HMC and adds a doubly-centered rater submodel that their formulation does not include. The interaction between hierarchical structure and NUTS (Hoffman et al., 2014) is well documented, with non-centered parameterizations (Paspiliopoulos et al., 2007; Betancourt and Girolami, 2015) essential for the geometry; we use a hierarchical local-scale prior on item loadings in the spirit of regularized continuous shrinkage (Pironen and Vehtari, 2017).

Rater effects in IRT are most commonly handled through the Facets framework (Linacre, 1989). Patz and Junker (1999) introduce the GLLTM, a 2PL/GPC generalization incorporating rater-by-item location shifts. Patz et al. (2002) demonstrate that Facets-style conditional-independence assumptions produce optimistically narrow standard errors

and propose the hierarchical rater model (HRM), in which ratings are conditionally dependent given an unobserved ideal score. The doubly-centered rater-by-item interaction we use is the generalizability-theory counterpart to this decomposition (Cronbach et al., 1972), orthogonalizing rater style from item difficulty by construction.

### 3 Model

Recovering how a student used the AI requires representing each student as a profile over several distinct behaviors rather than a single proficiency score. We therefore model usage with a multidimensional graded response model whose latent space is not committed in advance to any item-to-dimension mapping, and which must satisfy four requirements imposed by our setting: (i) it places each person in a  $D$ -dimensional latent space, so that qualitatively different usage profiles occupy different regions rather than different points on one axis; (ii) it accommodates a rubric mixing binary and four-point ordinal items in a single likelihood; (iii) it separates systematic rater style from item-level signal, since each document is rated by several coders; and (iv) it remains identifiable enough that the recovered dimensions can be interpreted. The remainder of this section specifies the likelihood, the loading prior, the person- and document-level structure, and the rater submodel in turn.

A multidimensional extension of the HRM for simple structured assessments (M-HRM) was introduced by Nieto and Casabianca (2019), with each item assigned to a single trait dimension via a confirmatory simple-structure specification. Our model differs from M-HRM in two design choices: (i) we identify the loading matrix post-hoc by rotation rather than confirmatorily, which suits the exploratory goal of surfacing AI-use dimensions without a pre-specified item-to-dimension mapping; and (ii) we model rater effects through additive severity and a doubly-centered Facets-style interaction rather than through the HRM ideal-rating layer. The HRM was developed in part to correct the optimistic standard errors that Facets-style conditional-independence assumptions produce (Patz et al., 2002); in our setting we rely on the full doubly-centered  $\xi_{rj}$  matrix to absorb structured rater disagreement, and our use of Facets-style rather than HRM machinery is a deliberate simplification rather than a claim of equivalence. Within the Facets-style choice, a rank-1 parameterization

$\xi_{rj} = \sigma_{\xi} s_r u_j$  proved unidentified (Appendix B), motivating the full doubly-centered matrix.

**Data and likelihood.** Observations are ordinal ratings  $y_n \in \{0, \dots, C_{j(n)} - 1\}$  on  $D$ -dimensional document traits, with  $C_j > 2$  for polytomous items and  $C_j = 2$  for binary items. Conditional on document trait  $\theta_d$ , item loading  $A_j$ , threshold  $\beta_{jc}$ , item noise  $\psi_j$ , rater severity  $\phi_r$ , and rater-by-item interaction  $\xi_{rj}$ , cumulative response probabilities follow the graded response form (Samejima, 1969)

$$P(y_n \geq c) = \sigma \left( \frac{A_j^\top \theta_d - \beta_{jc} - \phi_r - \xi_{rj}}{\psi_j} \right). \quad (1)$$

Thresholds are parameterized as  $\beta_{j,c} = \beta_j^{\text{anc}} + \sum_{c'=2}^c \delta_{j,c'}$  to enforce strict ordering. For binary items, only the first threshold is active and the cumulative-difference probabilities are renormalized over the two valid response categories; this differs from a looser parameterization in which non-existent threshold parameters are sampled but never enter the likelihood, and we found this distinction matters for PSIS-LOO reliability (Appendix F).

**Unconstrained loadings with item-level shrinkage.** Most in-model parameterizations we tried (Appendix B) either underidentified the model or distorted the posterior into local modes NUTS could not traverse. We therefore place a hierarchically shrunk prior on  $A$  with global scale  $\tau_A \sim \text{Half-Normal}(1)$ , per-item local scales  $\lambda_{A,j} \sim \text{Half-Normal}(1)$ , and  $A_{j,d} = \tau_A \lambda_{A,j} \tilde{A}_{j,d}$  with  $\tilde{A}_{j,d} \sim \mathcal{N}(0, 1)$ . The local scale accommodates the order-of-magnitude variation in item magnitudes our data exhibits. The unconstrained parameterization leaves  $A$  unidentified up to right-multiplication by an orthogonal matrix, per-column sign flips, and column permutations, with compensating transformations on  $\theta_d$ . All three symmetries are resolved post-hoc on posterior samples (Section 4).

**Person traits and document-level state.** Person traits  $\theta_p^{\text{pers}}$  are the quantity of substantive interest; the rest of the document-level model exists to estimate them free of the design’s confounds. Because each document is produced by one person working on one task under a specific reasoning context, topic, and domain-match status, we model the document trait additively as the person trait plus offsets for reasoning context, topic, and domain match, with a residual  $\epsilon_d$  for document-specific deviation. This decomposition keeps the person-level

latent interpretable as the student’s usage profile rather than a person-by-task blend, and it isolates the domain-match shift  $\delta^{\text{mat}}$  as its own term. We use a single Gaussian population prior on  $\theta^{\text{pers}}$  for parsimony; a Dirichlet-process mixture alternative (Appendix E) does not improve predictive density, so we retain the Gaussian.

#### Doubly-centered rater-by-item interaction.

Rater severity  $\phi_r$  is a non-centered scalar offset. The doubly-centered rater-by-item interaction is standard in generalizability theory (Cronbach et al., 1972) and many-facet Rasch models (Linacre, 1989); it separates rater style from item-level signal in the same spirit as the hierarchical rater model of Patz et al. (2002), though using a full-matrix rather than ideal-score parameterization (see Section 2 for the design choice rationale): we draw  $\tilde{\xi}_{rj} \sim \mathcal{N}(0, 1)$  and form

$$\xi_{rj} = \sigma_\xi \cdot (\tilde{\xi}_{rj} - \bar{\tilde{\xi}}_{r\cdot} - \bar{\tilde{\xi}}_{\cdot j} + \bar{\tilde{\xi}}_{\cdot\cdot}), \quad (2)$$

which is doubly centered to be orthogonal to  $\phi_r$  and  $\beta_j$  by construction. The effective parameter count is  $(R - 1)(J - 1)$ , well-supported for our  $N \approx 18,000$  observations.

## 4 Inference

**Sampler.** We fit with NUTS (Hoffman et al., 2014) in NumPyro (Phan et al., 2019; Bingham et al., 2019), eight chains with target acceptance 0.95, max tree depth 12, 3,000 warmup and 4,000 sampling iterations per chain.

**Post-hoc identification via MatchAlign.** The unconstrained loading matrix admits three symmetries: rotation  $A \rightarrow AR$ , per-column sign flips, and column permutations. Each chain finds the same posterior up to some combination of these transformations; chains are not in different posterior modes but in different oriented views of the same mode. Resolving symmetries on raw samples produces apparent non-convergence ( $\hat{R} \gg 1$ ) even though the underlying posterior is well-mixed.

We resolve all three symmetries post-hoc, adapting MatchAlign (Poworoznek et al., 2025) to the multidim-IRT setting. For each posterior draw we apply Varimax rotation to  $A$  to land on a sparsity-maximizing orientation. We designate a global pivot as the median-condition-number Varimax-rotated draw from the first valid chain (a chain is invalid if its posterior std on  $A$  collapses near zero). For every other Varimax-rotated draw we

find the optimal signed column permutation matching that draw to the pivot via the cost matrix  $M_{ij} = \min(\|A_{\cdot,i} - p_{\cdot,j}\|, \|A_{\cdot,i} + p_{\cdot,j}\|)$ , solved by the Hungarian algorithm (Kuhn, 1955); the choice of which norm achieves the minimum determines the sign for that pairing. The resulting rotation, permutation, and sign flips are applied jointly to all rotation-affected latent quantities ( $\theta^{\text{pers}}$ , condition effects, document residuals), preserving the likelihood by construction. Convergence diagnostics are evaluated on aligned samples for rotation-affected quantities and on raw samples for rotation-invariant quantities.

**Predictive evaluation.** We select the dimensionality  $D$  of the persons traits  $\theta_p^{\text{pers}}$  via PSIS-LOO (Vehtari et al., 2017), monitoring Pareto- $k$  as a reliability diagnostic for our model. Where bad- $k$  counts indicate LOO is unreliable for a given  $D$ , we report this and supplement with parameter recovery and loading-stability arguments.

## 5 Data

We collected ratings from  $P = 251$  German undergraduate students completing scenario-based tasks designed to assess Critical Online Reasoning (Zlatkin-Troitschanskaia et al., 2023). Each student completed multiple tasks crossing expertise level (binary) and reasoning context (binary) with task topic (8 academic topics), yielding 972 coded documents from 251 persons. Restricting the analysis to documents in which ChatGPT was actually used (the gating item 1.1) leaves  $D_{\text{docs}} = 750$  documents from  $P = 191$  persons, which form the analytic sample for all models below. The original coding schema comprised 15 items; Table 1 reports per-item Krippendorff’s  $\alpha$  for the 13 with enough observations to estimate it (2.2.1 and 2.2.2 are omitted). The analysis uses the  $J = 9$ -item subset defined in Appendix H.

## 6 Results

**Dimension selection.** We separate two questions: whether the latent space is multidimensional, and how many dimensions to retain. PSIS-LOO (Table 2) answers the first decisively. It is reliable for  $D \in \{1, 2, 3\}$  and increases monotonically ( $D=1 \rightarrow 2$ : +1,081 ELPD;  $D=2 \rightarrow 3$ : +755), rejecting unidimensionality and selecting the largest dimension it can reliably estimate,  $D = 3$  (28 bad- $k$  of 18,444). It cannot answer the second: at  $D = 4$ , 854 of 18,444 observations exceed

| Item  | Alias | Description                                                                                     | $\alpha$ |
|-------|-------|-------------------------------------------------------------------------------------------------|----------|
| 1.1   |       | ChatGPT was used at least once while working on the current page (gate)                         | .96      |
| 2.1.1 |       | Person enters a task related prompt                                                             | .92      |
| 2.1.2 | 15    | Person copies (almost) the entire task description and uses it as a prompt                      | .97      |
| 2.1.3 |       | Person copies part of the task description and uses it as a prompt                              | .31      |
| 2.1.4 | 17    | Person asks specific, self-formulated questions (to work on the task directly or similar)       | .33      |
| 2.1.5 | 18    | Person asks general/abstract questions (to gain a better understanding of the topic or similar) | .40      |
| 2.2.3 |       | Person instructs ChatGPT to formulate a text based on given text                                | .00      |
| 2.2.4 | 16    | Person instructs ChatGPT to revise a previous given answer                                      | .48      |
| 2.3.1 | 14    | Person enters several different prompts to work on different tasks                              | .65      |
| 2.3.2 | 19    | Person enters several different prompts to work on different aspects of one task                | .41      |
| 3.1   | 12    | ChatGPT was used in the Citation Sources                                                        | .65      |
| 3.2.1 | 13    | ChatGPT was used in the Position statement                                                      | .91      |
| 3.3.1 | 11    | ChatGPT was used in the Decision Criteria                                                       | .85      |

Table 1: Per-item Krippendorff’s  $\alpha$  computed across 972 coded units. Items 2.2.1 and 2.2.2 are not shown due to insufficient observation count. Following Krippendorff (2004),  $\alpha \geq .80$  supports definitive conclusions and  $\alpha \geq .667$  tentative conclusions.

$k > 0.7$  with non-finite contributions, violating the conditions for calibrated comparison (Vehtari et al., 2017), so any apparent in-sample improvement at  $D = 4$  carries no selection weight. We resolve  $D = 3$  versus  $D = 4$  on grounds LOO leaves open: at  $D = 4$  the added dimension is dominated by a single item (Appendix D), over-specifying rather than refining the structure, and parameter recovery degrades in loadings, person traits, and most sharply the match shift (sign flips on three of ten runs versus one of ten at  $D = 3$ ; Appendix F). We take  $D = 3$  as primary and report  $D = 4$  as a sensitivity check.

| $D$ | ELPD   | SE | $p_{100}$ | $k > 0.7$ |
|-----|--------|----|-----------|-----------|
| 1   | −10120 | 89 | 664       | 52        |
| 2   | −9039  | 92 | 1157      | 74        |
| 3   | −8284  | 93 | 1550      | 28        |
| 4   | —      | —  | —         | 854       |

Table 2: PSIS-LOO across candidate dimensions for the multidim GRM. PSIS-LOO is reliable at  $D \leq 3$ . At  $D = 4$ , 854 of 18,444 observations exceed the  $k > 0.7$  threshold and individual log predictive contributions are non-finite, so ELPD, SE, and  $p_{100}$  are undefined. The bad- $k$  count is the operative reliability summary at  $D = 4$ .

**Convergence.** All eight chains pass the chain-failure diagnostic at  $D = 3$ , and pairwise correlations of chain-mean  $A$  equal 1.000 at three decimal places across all 28 chain pairs after MatchAlign. The rotation-affected derived quantities ( $\theta^{\text{pers}}$ ,  $\delta_{\text{match}}$ ,  $\delta_{\text{topic}}$ ) all satisfy  $\hat{R} < 1.01$  on

aligned samples (Vehtari et al., 2021; Gelman and Rubin, 1992).

**Loading structure.** Figure 1 shows the aligned loading matrix at  $D = 3$ . Column norms are  $[2.99, 5.97, 2.77]$ , with  $D_2$  accounting for the largest share of variance. The three dimensions admit a clean substantive reading.  $D_1$  collects the three binary prompting items 7 (concrete questions), 8 (abstract questions), and 9 (multi-prompt task aspects), each with dominance fraction (defined as  $|A_{j,d^*}|/\|A_{j,\cdot}\|_2$  for dominant dimension  $d^*$ ) at or above 0.94. We read  $D_1$  as *active prompting effort*: a dimension on which students who probe the AI with formulated questions and structured aspect-prompts score high.  $D_2$  collects polytomous items 1 and 3 (decision criteria delivered by AI; position statements delivered by AI) along with binary items 4 (multi-task prompting), 5 (full task description as prompt), and 6 (revision request), with dominance fractions between 0.81 and 0.99. We read  $D_2$  as *AI delegation of substantive content*: students score high when the AI is doing the thinking and they are receiving the output.  $D_3$  has its highest loading on polytomous item 2 (citation sources delivered by AI), with dominance fraction 0.82; we read  $D_3$  as *AI-delivered citation use*, capturing the specific case of relying on AI for sourcing rather than for substantive reasoning. The clean simple structure has one exception: item 4 (multi-task prompting) loads credibly on all three dimensions: its  $D_2$  loading dominates (dominance fraction 0.81), but it carries non-negligible

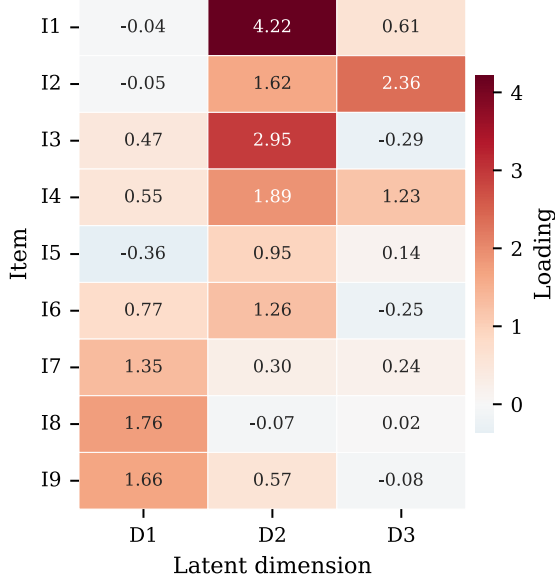


Figure 1: Posterior mean of loading matrix  $A$  at  $D = 3$  after Varimax rotation and MatchAlign. Items 7–9 (binary prompting items) commit to  $D_1$  with dominance fractions above 0.94; items 1, 3 and binary items 4–6 load primarily on  $D_2$ ; item 2 (citation sources) loads primarily on  $D_3$ . After applying the polarity convention, the dominant loading in each column is positive.

raw loadings on  $D_1$  (0.55) and  $D_3$  (1.23) as well. This is substantively plausible, since issuing distinct prompts for distinct tasks is partly an effort behavior, partly a delegation behavior, and partly co-occurs with sourcing; we treat item 4 as a deliberate cross-loader rather than as evidence against the three-dimensional structure. Per-item loadings, posterior stds, and dominance fractions are in Appendix I.

**Stability of loadings across  $D$ .** The loading structure is stable across choices of  $D$  in a way that supports the substantive reading. The prompting items (7–9) consistently form a dimension distinct from the AI-contribution items (1–3) at every  $D \in \{3, 4\}$ . At  $D = 4$  the structure refines rather than reorganizes. The polytomous citation-source item (item 2) already separates onto its own dimension at  $D = 3$ ; the only further change at  $D = 4$  is that binary item 5 (full task description as prompt) splits into a one-item dimension, while delegation items 1 and 3 remain together. We treat this one-item dimension as evidence that  $D = 4$  over-specifies the latent space, separating an item that does not warrant its own dimension on substantive grounds. Appendix D reports the  $D = 4$  loadings in full.

**Comparison against a Dirichlet-process variant.** We additionally fit a variant in which the single Gaussian on  $\theta^{\text{pers}}$  is replaced with a truncated Dirichlet-process mixture of Gaussians (DP-MIRT) at matched  $D$ . At  $D = 3$  the two specifications produce essentially identical predictive performance: ELPD  $-8275$  for DP-MIRT and  $-8284$  for the GRM, well within one standard error. PSIS-LOO bad- $k$  counts are 24 (DP-MIRT) versus 28 (GRM) at  $D = 3$ . We report DP-MIRT at  $D = 3$  as a sensitivity check in Appendix E.

**Two illustrative students.** The clearest contrast lies along  $D_1$  (prompting effort) versus  $D_2$  (AI delegation). Student A (person 55) has posterior-mean traits  $\theta_{55}^{\text{pers}} = (+1.11, -3.57, +0.91)$ : high on prompting effort, strongly low on AI-delivered substance, and positive on citation reliance. Student B (person 118) shows the opposite pattern on the prompting–delegation axis,  $\theta_{118}^{\text{pers}} = (-1.38, +2.02, +0.10)$ , with  $D_1$  and  $D_2$  both clearly resolved away from zero ( $D_1 \in [-2.26, -0.52]$ ,  $D_2 \in [+1.09, +2.95]$ ) while  $D_3$  is indistinguishable from zero. Student A probes the AI with formulated questions and revision requests, writes their own positions and decision criteria, and relies on the AI for sourcing—a research-assistant pattern. Student B does little prompt work and receives extensive substantive content from the AI without elevated citation use. These two students make qualitatively different use of the same tool; aggregate scoring would conflate them.

## 7 Discussion

**What the dimensions mean.** The three latent dimensions admit a coherent reading consistent with the rubric structure.  $D_1$  captures *active prompting effort*: the use of formulated questions, structured aspect-prompts, and distinct prompts for distinct purposes. The dimension maps onto the qualitative distinction between copy-and-paste prompting, single reformulated questions, and multi-question interaction modes documented in log-based studies of student ChatGPT use (Sawalha et al., 2024; Stojanov et al., 2024; Güner and Er, 2025); the present model provides a probabilistic, multidimensional measurement of the same underlying heterogeneity rather than a cluster assignment.  $D_2$  captures *AI delegation of substantive content*: the degree to which decision criteria, position statements, and structured task framing are produced by the AI rather than the student.  $D_3$  captures *AI-delivered*

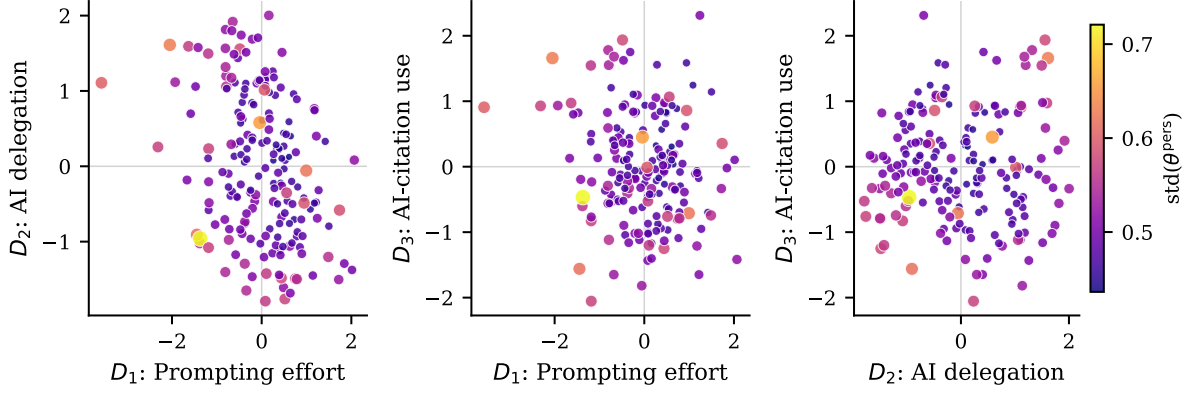


Figure 2: Person latent traits  $\theta_p^{\text{pers}}$  at  $D = 3$ , displayed as three pairwise projections of the latent space. Each point is one of the  $P = 191$  students, placed at the posterior mean of their trait vector; color encodes posterior standard deviation  $\text{std}(\theta^{\text{pers}})$ , with yellow indicating higher estimation uncertainty.

*citation use*: the specific case of relying on the AI for sourcing rather than for reasoning. The separation of  $D_3$  from  $D_2$  is substantively informative: citations are factual rather than substantive, and a student can rely on AI for sources while still writing their own positions, or vice versa. The model treats these as distinct uses with different rubric correlates.

**Domain match and the research-assistant pattern.** The paired domain-match findings on  $D_1$ ,  $D_2$ , and  $D_3$  admit a substantive reading. Among students whose major is represented in the task topics, working in one’s own academic domain shifts them toward more prompting effort ( $+0.29$ ,  $P(+) = 1.00$ ) and more citation-seeking behaviour ( $+0.28$ ,  $P(+) = 0.99$ ), and away from delegated substantive content ( $-0.13$ ,  $P(+) = 0.08$ ). This is consistent with a “research assistant” pattern of in-domain AI use: students with subject-matter familiarity treat the AI as a tool for sourcing and probing while retaining authorship of the substantive content. The opposite pattern at low domain match (less prompting, less citation-seeking, more delegated content) suggests that students without domain expertise lean more heavily on AI-produced substance. Whether this reflects strategic choice (students know what they don’t know and outsource accordingly) or competence-driven defaults (students who can prompt well do, others fall back on whatever the AI produces) cannot be distinguished from cross-sectional data.

## 8 Implications and Limitations

Two components of our methodology may generalize. Post-hoc MatchAlign offers an alternative to in-model loading constraints when the constraints distort the likelihood; we documented three in-model attempts that fail (Appendix B) before adopting the post-hoc route. The doubly-centered rater-by-item interaction suits rating designs where rank-restricted forms are unidentified.

Our analysis is based on a single dataset from a single educational context, so the substantive reading of dimensions and effects should not be generalized without replication. On synthetic data the rater submodel recovers consistently (mean  $\phi_r$  and  $\xi_{r,j}$  correlations of at least 0.97), and the loading structure recovers well at  $D = 3$  but degrades at  $D = 4$ ; full recovery metrics are in Table 3. Six of the nine retained items have Krippendorff’s  $\alpha$  below 0.667. The model accommodates this through item-level noise and the rater-by-item interaction, and the posterior  $\psi_j$  values for these items are not elevated relative to higher- $\alpha$  items, suggesting much of what  $\alpha$  records is structured rater style rather than random measurement error. A direct check (Appendix G) that refits on only the five items at  $\alpha > 0.5$  shows the delegation–citation separation and the in-domain citation effect persist without the sub-threshold items; the active-prompting dimension, however, is anchored entirely by sub-threshold items and cannot be corroborated this way, so it rests on the  $\psi_j$  argument alone. We retain the items because their substantive content is central to the multidimensional structure we examine, but readers should treat the dimensions that depend on them as partially reconstructed from rater-

specific interpretations of the rubric: the model can fit around rater disagreement, but it cannot generate construct validity that the underlying coding does not support.

Future work includes longitudinal extensions, applications to larger samples that would let dimensionality selection rest more heavily on PSIS-LOO at higher  $D$ , and a rubric revision for the lower- $\alpha$  items grounded in the dimensional structure we identify.

## Acknowledgements

This work was supported by the research group CORE (FOR 5404, project number 462702138), funded by the German Research Foundation (DFG).

## References

- Tawfiq Ammari, Meilun Chen, S M Mehedi Zaman, and Kiran Garimella. 2026. [How students \(really\) use chatgpt: Uncovering experiences among undergraduate students](#). *Preprint*, arXiv:2505.24126.
- Christian Alßmann, Jens Boysen-Hogrefe, and Markus Pape. 2016. [Bayesian analysis of static and dynamic factor models: An ex-post approach towards the rotation problem](#). *Journal of Econometrics*, 192(1):190–206.
- Anton A Béguin and Ceec AW Glas. 2001. [Mcmc estimation and some model-fit analysis of multidimensional irt models](#). *Psychometrika*, 66(4):541–561.
- Michael Betancourt and Mark Girolami. 2015. [Hamiltonian Monte Carlo for Hierarchical Models](#). In Satyanshu K. Upadhyay, Umesh Singh, Dipak K. Dey, and Appaia Loganathan, editors, *Current Trends in Bayesian Methodology with Applications*, 0 edition, pages 119–142. Chapman and Hall/CRC.
- A. Bhattacharya and D. B. Dunson. 2011. [Sparse bayesian infinite factor models](#). *Biometrika*, 98(2):291–306.
- Eli Bingham, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul A. Szerlip, Paul Horsfall, and Noah D. Goodman. 2019. [Pyro: Deep universal probabilistic programming](#). *J. Mach. Learn. Res.*, 20:28:1–28:6.
- Cecilia Ka Yuk Chan. 2023. [A comprehensive ai policy education framework for university teaching and learning](#). *International journal of educational technology in higher education*, 20(1):38.
- Jacob Cohen. 1960. [A coefficient of agreement for nominal scales](#). *Educational and psychological measurement*, 20(1):37–46.
- Lee J. Cronbach, Goldine C. Gleser, Harinder Nanda, and Nageswari Rajaratnam. 1972. *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. Wiley, New York.
- Phillip Dawson, Margaret Bearman, Mollie Dollinger, and David Boud. 2024. [Validity matters more than cheating](#). *Assessment & Evaluation in Higher Education*, 49(7):1005–1016.
- Ying Dong, Wei Xu, Jiayan Huang, and Kerr Yann. 2025. [Validating and refining a multi-dimensional scale for measuring ai literacy in education using the rasch model](#). *Humanities and Social Sciences Communications*, 12(1):1317.
- Thomas S. Ferguson. 1973. [A Bayesian Analysis of Some Nonparametric Problems](#). *The Annals of Statistics*, 1(2).
- Sylvia Frühwirth-Schnatter, Darjus Hosszejni, and Hedibert Freitas Lopes. 2025. [Sparse Bayesian Factor Analysis When the Number of Factors Is Unknown \(with Discussion\)](#). *Bayesian Analysis*, 20(1):213 – 344.
- Andrew Gelman and Donald B Rubin. 1992. [Inference from iterative simulation using multiple sequences](#). *Statistical science*, 7(4):457–472.
- John Geweke and Guofu Zhou. 1996. [Measuring the pricing error of the arbitrage pricing theory](#). *The Review of Financial Studies*, 9(2):557–587.
- Hacer Güner and Erkan Er. 2025. [Ai in the classroom: Exploring students’ interaction with chatgpt in programming learning](#). *Education and Information Technologies*, 30(9):12681–12707.
- Neda Hadinejad, Katarina Sperling, and Cormac McGrath. 2025. [Generative ai chatbots in higher education: Student experiences and perceived ethical challenges](#). *Computers and Education Open*, 9:100311.
- Matthew D Hoffman, Andrew Gelman, and 1 others. 2014. [The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo](#). *J. Mach. Learn. Res.*, 15(1):1593–1623.
- Hemant Ishwaran and Lancelot F James. 2001. [Gibbs sampling methods for stick-breaking priors](#). *Journal of the American Statistical Association*, 96(453):161–173.
- Henry F Kaiser. 1958. [The varimax criterion for analytic rotation in factor analysis](#). *Psychometrika*, 23(3):187–200.
- Leonora Kaldaras, Hope O. Akaeze, and Mark D. Reckase. 2024. [Developing valid assessments in the era of generative artificial intelligence](#). *Frontiers in Education*, 9.
- Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke

- Hüllermeier, and 1 others. 2023. [Chatgpt for good? on opportunities and challenges of large language models for education](#). *Learning and individual differences*, 103:102274.
- Jinhee Kim, Sang-Soog Lee, Rita Detrick, Jialin Wang, and Na Li. 2025. [Students-generative ai interaction patterns and its impact on academic writing](#). *Journal of Computing in Higher Education*, 38:504–525.
- Carmen Köhler and Johannes Hartig. 2024. [Chatgpt in higher education: Measurement instruments to assess student knowledge, usage, and attitude](#). *Contemporary educational technology*, 16(4):ep528.
- Klaus Krippendorff. 2004. [Reliability in content analysis: Some common misconceptions and recommendations](#). *Human communication research*, 30(3):411–433.
- Harold W Kuhn. 1955. [The hungarian method for the assignment problem](#). *Naval research logistics quarterly*, 2(1-2):83–97.
- Jiguang Li, Robert Gibbons, and Veronika Ročková. 2025. [Sparse bayesian multidimensional item response theory](#). *Journal of the American Statistical Association*, 120(552):2592–2605.
- John Michael Linacre. 1989. *Many-Facet Rasch Measurement*. MESA Press, Chicago.
- David J. C. MacKay. 1992. [Bayesian interpolation](#). *Neural Computation*, 4(3):415–447.
- Hunter McNichols, Fareya Ikram, and Andrew Lan. 2026. [The studychat dataset: Analyzing student dialogues with chatgpt in an artificial intelligence course](#). In *Proceedings of the LAK26: 16th International Learning Analytics and Knowledge Conference*, pages 53–63.
- Radford M. Neal. 1996. *Bayesian Learning for Neural Networks*. Number 118 in Lecture Notes in Statistics. Springer, New York.
- Ricardo Nieto and Jodi M. Casabianca. 2019. [Accounting for rater effects with the hierarchical rater model framework when scoring simple structured constructed response tests](#). *Journal of Educational Measurement*, 56(3):547–581.
- Omiros Papaspiliopoulos, Gareth O. Roberts, and Martin Sköld. 2007. [A general framework for the parametrization of hierarchical models](#). *Statistical Science*, 22(1):59–73.
- Panagiotis Papastamoulis and Ioannis Ntzoufras. 2022. [On the identifiability of bayesian factor analytic models](#). *Statistics and Computing*, 32(2):23.
- Richard J. Patz and Brian W. Junker. 1999. [Applications and extensions of mcmc in irt: Multiple item types, missing data, and rated responses](#). *Journal of Educational and Behavioral Statistics*, 24(4):342–366.
- Richard J. Patz, Brian W. Junker, Matthew S. Johnson, and Louis T. Mariano. 2002. [The hierarchical rater model for rated test items and its application to large-scale educational assessment data](#). *Journal of Educational and Behavioral Statistics*, 27(4):341–384.
- Mike Perkins, Leon Furze, Jasper Roe, and Jason MacVaugh. 2024. [The Artificial Intelligence Assessment Scale \(AIAS\): A framework for ethical integration of Generative AI in Educational Assessment](#). *Journal of University Teaching and Learning Practice*, 21(6):49–66.
- Du Phan, Neeraj Pradhan, and Martin Jankowiak. 2019. [Composable effects for flexible and accelerated probabilistic programming in numpyro](#). *arXiv preprint arXiv:1912.11554*.
- Juho Piironen and Aki Vehtari. 2017. [Sparsity information and regularization in the horseshoe and other shrinkage priors](#). *Electronic Journal of Statistics*, 11(2):5018 – 5051.
- Evan Poworoznek, Niccolo Anceschi, Federico Ferrari, and David Dunson. 2025. [Efficiently resolving rotational ambiguity in bayesian matrix sampling with matching](#). *Bayesian Analysis*, 1(1):1–22.
- Mark D. Reckase. 2009. *Unidimensional Item Response Theory Models*, pages 11–55. Springer New York, New York, NY.
- Veronika Ročková and Edward I. George. 2016. [Fast bayesian factor analysis via automatic rotations to sparsity](#). *Journal of the American Statistical Association*, 111(516):1608–1622.
- Fumiko Samejima. 1969. [Estimation of latent ability using a response pattern of graded scores](#). *Psychometrika*, 34(S1):1–97.
- Ghadeer Sawalha, Imran Taj, and Abdulhadi Shoufan. 2024. [Analyzing student prompts and their effect on chatgpt’s performance](#). *Cogent Education*, 11(1):2397200.
- Jayaram Sethuraman. 1994. [A constructive definition of dirichlet priors](#). *Statistica Sinica*, 4(2):639–650.
- Matthew Stephens. 2000. [Dealing with label switching in mixture models](#). *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 62(4):795–809.
- Ana Stojanov, Qian Liu, and Joyce Hwee Ling Koh. 2024. [University students’ self-reported reliance on chatgpt for learning: A latent profile analysis](#). *Computers and Education: Artificial Intelligence*, 6:100243.
- Aki Vehtari, Andrew Gelman, and Jonah Gabry. 2017. [Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC](#). *Statistics and Computing*, 27(5):1413–1432.

Aki Vehtari, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian Bürkner. 2021. [Rank-normalization, folding, and localization: An improved  \$\hat{R}\$  for assessing convergence of mcmc \(with discussion\)](#). *Bayesian analysis*, 16(2):667–718.

Xingyao Xiao, Sophia Rabe-Hesketh, and Anders Skroindal. 2025. [Bayesian identification and estimation of growth mixture models](#). *Psychometrika*, 90(2):442–475.

Da-Wei Zhang, Jia Yue Tan, Yu Yang Chew, Lisha Hew, and Jia Yee Choo. 2025. [Development and validation of the generative ai engagement scale](#). *Computers in Human Behavior: Artificial Humans*, 6:100221.

Olga Zlatkin-Troitschanskaia, Dimitri Molerov, Jennifer Fischer, and Johannes Hartig. 2023. [Research unit critical online reasoning in higher education \(for core\) framework, research design, and structure](#). *CORE’working papers*.

## A Full Model Specification

We provide the complete prior specification for reference. Let  $D$  denote the latent dimension,  $J$  the number of items,  $R$  the number of raters,  $P$  the number of persons.

### Item parameters.

$$\tau_A \sim \text{Half-Normal}(1) \quad (3)$$

$$\lambda_{A,j} \sim \text{Half-Normal}(1), \quad j = 1, \dots, J \quad (4)$$

$$\tilde{A}_{j,d} \sim \mathcal{N}(0, 1), \quad d = 1, \dots, D \quad (5)$$

$$A_{j,d} = \tau_A \lambda_{A,j} \tilde{A}_{j,d} \quad (6)$$

$$\psi_j \sim \text{LogNormal}(0, 0.2) \quad (7)$$

$$\beta_j^{\text{anc}} \sim \mathcal{N}(0, 2) \quad (8)$$

$$\delta_{j,c} \sim \text{LogNormal}(-0.3, 0.5) \quad (9)$$

Item thresholds are reconstructed as  $\beta_{j,1} = \beta_j^{\text{anc}}$  and  $\beta_{j,c} = \beta_j^{\text{anc}} + \sum_{c'=2}^c \delta_{j,c'}$  for  $c > 1$ , ensuring strict ordering. For binary items, only the first threshold enters the likelihood, with cumulative-difference probabilities renormalized over the two valid response categories.

### Rater parameters.

$$\sigma_\phi \sim \text{Half-Normal}(1) \quad (10)$$

$$\tilde{\phi}_r \sim \mathcal{N}(0, 1), \quad r = 1, \dots, R \quad (11)$$

$$\phi_r = \sigma_\phi \tilde{\phi}_r \quad (12)$$

$$\sigma_\xi \sim \text{Half-Normal}(0.3) \quad (13)$$

$$\tilde{\xi}_{rj} \sim \mathcal{N}(0, 1) \quad (14)$$

$$\xi_{rj} = \sigma_\xi (\tilde{\xi}_{rj} - \bar{\xi}_{r\cdot} - \bar{\xi}_{\cdot j} + \bar{\xi}_{\cdot\cdot}) \quad (15)$$

The doubly-centered  $\xi_{rj}$  satisfies  $\sum_r \xi_{rj} = 0$  and  $\sum_j \xi_{rj} = 0$ , making it orthogonal to  $\phi_r$  and  $\beta_j$  by construction.

### Person traits.

$$\theta_p^{\text{pers}} \sim \mathcal{N}(0, I_D), \quad p = 1, \dots, P \quad (16)$$

### Document state and condition effects.

$$\sigma_l \sim \text{Half-Normal}(0.5) \quad (17)$$

$$\sigma_t \sim \text{Half-Normal}(0.2) \quad (18)$$

$$\sigma_\epsilon \sim \text{Half-Normal}(0.3) \quad (19)$$

$$\tilde{\delta}_\ell^{\text{var}} \sim \mathcal{N}(0, I_D), \quad \delta_\ell^{\text{ctx}} = \sigma_l \tilde{\delta}_\ell^{\text{var}} \quad (20)$$

$$\tilde{\delta}_t^{\text{top}} \sim \mathcal{N}(0, I_D), \quad \delta_t^{\text{top}} = \sigma_t \tilde{\delta}_t^{\text{top}} \quad (21)$$

$$\delta^{\text{mat}} \sim \mathcal{N}(0, I_D) \quad (22)$$

$$\tilde{\epsilon}_d \sim \mathcal{N}(0, I_D), \quad \epsilon_d = \sigma_\epsilon \tilde{\epsilon}_d \quad (23)$$

$$\theta_d = \theta_{p(d)}^{\text{pers}} + \delta_{\ell(d)}^{\text{ctx}} + \delta_{t(d)}^{\text{top}} + \delta^{\text{mat}} m_d + \epsilon_d \quad (24)$$

**Likelihood.** For observation  $n$  rating document  $d$  on item  $j$  by rater  $r$ ,

$$\eta_{njc} = A_j^\top \theta_d - \beta_{j,c} - \phi_r - \xi_{rj} \quad (25)$$

$$P(y_n \geq c) = \sigma(\eta_{njc}/\psi_j) \quad (26)$$

$$P(y_n = c) = P(y_n \geq c) - P(y_n \geq c + 1) \quad (27)$$

## B Negative Results: In-Model Identification and Match-Effect Variants

We document four model variants we attempted and rejected. The first three address loading-matrix identification and predate our adoption of post-hoc MatchAlign; the fourth tests whether domain-match effects vary with a person’s position in the latent space.

**Lower-triangular anchor.** Following [Geweke and Zhou \(1996\)](#), we constrained the first  $D$  polytomous items to form an anchor block  $A^{\text{anc}} \in \mathbb{R}^{D \times D}$  with  $A_{ij}^{\text{anc}} = 0$  for  $j > i$  and  $A_{ii}^{\text{anc}} > 0$ . This fixes rotation up to reflection and is the standard solution. We observed a residual scale-swap multimodality on the diagonal: chains found configurations with  $A_{11}^{\text{anc}} \gg A_{22}^{\text{anc}}$  and configurations with a more balanced diagonal that explained the data equally well, with NUTS unable to traverse between them.

**Ordered positive diagonal.** We extended the lower-triangular anchor with a strict ordering constraint on the diagonal:  $A_{11}^{\text{anc}} > A_{22}^{\text{anc}} > \dots > A_{DD}^{\text{anc}} > 0$ , implemented via a cumulative-softplus construction on the log-diagonal following

Frühwirth-Schnatter et al. (2025). This eliminated the scale-swap multimodality but introduced its own pathology: chains struggled to traverse the ordering boundary during warmup, leading to occasional single-chain failures on the weakest diagonal entry.

**Factorized  $A = WL \text{diag}(s)$  with ARD ordering.** Attempting to remove hard zeros in favor of soft sparsity, we factorized the loading matrix into orthogonal  $W$ , lower-triangular  $L$  with unit diagonal, and ARD-style (MacKay, 1992; Neal, 1996) scale vector  $s$  ordered by  $s_1 > s_2 > \dots > s_D$ . The ARD scaling was intended to break column permutation symmetry by sorting columns in decreasing variance contribution. This failed: different chains found different item-to-dimension assignments. Our interpretation is that the ARD ordering breaks one symmetry but the rotation symmetry of  $W$  remains, and the two interact in a way that NUTS cannot navigate.

**Latent-axis-conditional domain match.** We hypothesized that the domain-match effect might depend on a person’s position in the latent trait space, with different student types responding differently to working in their academic domain. We tested this by replacing the global  $\delta^{\text{mat}} m_d$  with  $(\delta^{\text{baseline}} + M\theta_p^{\text{pers}})m_d$ , where  $M \in \mathbb{R}^{D \times D}$  captures linear interaction between trait position and match shift, with prior  $M_{ij} \sim \mathcal{N}(0, \tau_M^2)$  and  $\tau_M \sim \text{HalfNormal}(0.3)$ . The fit at  $D = 4$  produced a posterior mean of  $M$  with diagonal entries between  $-0.04$  and  $-0.05$  and off-diagonal entries all below  $0.01$  in absolute value, consistent with  $M \approx -0.045 I$ . This corresponds to a uniform 4.5% shrinkage of all trait dimensions under match condition rather than direction-specific interaction. We read this as the data ruling out trait-dependent match effects in favor of the simpler global shift in the main specification. The fit also showed identifiability tension between  $\delta^{\text{baseline}}$  and  $M$  ( $\hat{R} > 1.3$  on the baseline), further suggesting the interaction was absorbing variance the global shift handles cleanly.

**Lessons.** In every case, we observed that adding constraints during sampling tightened parts of the posterior geometry while opening new pathologies elsewhere. This is consistent with the broader intuition that in-model identification constraints distort the likelihood and create new local modes (Papastamoulis and Ntzoufras, 2022; Poworoznek

et al., 2025); it parallels the analogous finding for mixture-model label switching, where Stephens (2000) shows that artificial identifiability constraints on mixture components fail in general and post-hoc relabelling is preferable. The post-hoc approach succeeds because it lets the sampler use the full likelihood and resolves identification only on samples after sampling completes.

## C MatchAlign Algorithm and Chain Diagnostic

**MatchAlign procedure.** Algorithm 1 gives the full pipeline.

---

### Algorithm 1 MatchAlign for multidim IRT.

---

**Input:** per-chain samples  $\{A^{(c,t)}\}_{c,t}$  and rotation-affected latent quantities.

**Output:** aligned samples in a globally fixed orientation.

1. Detect collapsed chains: exclude any chain  $c$  for which  $\text{mean}_t \text{std}_t A^{(c)} < 5 \times 10^{-2}$ .
2. For every  $(c, t)$  in the surviving chains, apply Varimax rotation to  $A^{(c,t)}$ , recording rotation matrices  $R^{(c,t)}$ .
3. Choose a global pivot  $p$ : the Varimax-rotated draw with median condition number from the first surviving chain.
4. For each  $(c, t)$ , form the cost matrix

$$M_{ij} = \min(\|A_{:,i}^{(c,t)} - p_{:,j}\|, \|A_{:,i}^{(c,t)} + p_{:,j}\|)$$

and recover the sign  $s_{ij} \in \{-1, +1\}$  achieving the minimum. Solve the assignment  $\pi^{(c,t)}$  by the Hungarian algorithm (Kuhn, 1955); the corresponding signs  $s^{(c,t)}$  follow.

5. Apply  $R^{(c,t)}$ ,  $\pi^{(c,t)}$ ,  $s^{(c,t)}$  to  $A^{(c,t)}$  and to all rotation-affected latent quantities ( $\theta^{\text{pers}}$ , condition effects  $\delta^{\text{ctx}}$ ,  $\delta^{\text{top}}$ ,  $\delta^{\text{mat}}$ , document residuals  $\epsilon_d$ ).
- 

The pivot selection by median condition number rather than maximum sparsity is deliberate: a high-sparsity Varimax solution can have nearly-zero columns that produce ill-conditioned matching, while the median condition number gives a representative anchor that is well-conditioned for assignment.

## D $D = 4$ Sensitivity Check

We fit the model at  $D = 4$  as a sensitivity check on dimensionality selection. The fit converges (all eight chains pass the chain-failure diagnostic; pairwise correlations of chain-mean  $A$  equal 1.000 after MatchAlign), but PSIS-LOO is unreliable at  $D = 4$  with 854 of 18,444 observations exceeding

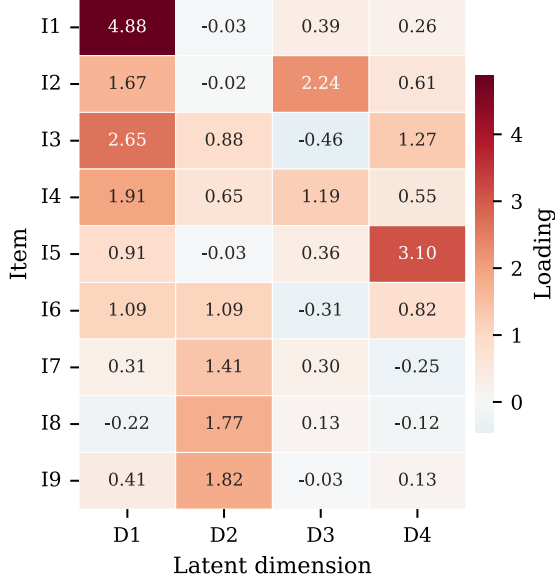


Figure 3: Posterior mean of loading matrix  $A$  at  $D = 4$  after Varimax rotation and MatchAlign. The structure refines rather than reorganises the  $D = 3$  solution.  $D_1$  is anchored by items 1 (decision criteria, 4.88) and 3 (position statement, 2.65), corresponding to AI delegation of substantive content;  $D_2$  collects the three binary prompting items 7–9 (concrete questions, abstract questions, multi-aspect prompting), corresponding to active prompting effort. The dimension labels therefore exchange positions relative to the  $D = 3$  solution, a MatchAlign rotation artefact with no substantive consequence.  $D_3$  retains item 2 (source citations, 2.24) as the AI-citation use dimension, consistent with  $D = 3$ .  $D_4$  is dominated by a single item: item 5 (full task description as prompt, 3.10), which was absorbed into the delegation dimension at  $D = 3$ . The emergence of a one-item dimension at  $D = 4$  is the primary evidence that this solution over-specifies the latent space; item 5 does not warrant a dedicated dimension on substantive grounds.

$k > 0.7$  and individual log predictive contributions producing non-finite values. The loading structure refines rather than reorganizes the  $D = 3$  structure: prompting items 7–9 still anchor a single dimension; the citation-source item (item 2) already split off at  $D = 3$ ; delegation items 1 and 3 remain together on one dimension; and item 5 (full task description as prompt) forms a one-item dimension. We treat the latter as evidence that  $D = 4$  over-specifies the latent space, separating an item that does not warrant its own dimension on substantive grounds. Per-item loadings at  $D = 4$  are reported in Figure 3.

## E DP-MIRT Sensitivity Check

We additionally fit the model with the single Gaussian on  $\theta^{\text{pers}}$  replaced by a truncated DP mixture of Gaussians (Ferguson, 1973; Sethuraman, 1994) with up to  $K_{\text{max}} = 12$  components, using the stick-breaking parameterization through the inverse Beta CDF (Ishwaran and James, 2001). Identification of cluster-component labels is resolved by sorting cluster centroids on their first dimension within each MatchAlign-aligned draw. At  $D = 3$ , the DP fit produces ELPD  $-8275$  (SE 93) with 24 of 18,444 bad- $k$  observations, compared to ELPD  $-8284$  with 28 bad- $k$  for the GRM. The DP variant improves PSIS-LOO reliability most likely by accommodating extreme persons in dedicated components, but does not improve predictive density at the population level. The loading structure, rater submodel, and domain-match effects under the DP variant coincide with the GRM specification within posterior uncertainty. We retain the GRM as the primary specification on parsimony grounds.

## F Parameter Recovery

We evaluate parameter recovery on synthetic data generated from the fitted model. For each combination  $(D, P) \in \{3, 4\} \times \{100, 200\}$ , we fix  $D$ -independent parameters (rater severities  $\phi_r$ , rater-by-item interactions  $\xi_{r,j}$ , item thresholds  $\beta_{j,c}$ , item noise  $\psi_j$ ) at posterior means from the real-data fit, and synthesize fresh  $D$ -dependent parameters: loadings  $A$  at the posterior-mean scale, person traits drawn from  $\mathcal{N}(0, I_D)$ , and condition effects drawn from the model priors. We then draw ratings under the GRM likelihood and refit the full model with the same sampler settings as the main fit, applying MatchAlign to recover identified parameters. We run five seeds per condition, for 20 synthetic datasets total.

**Recovery metrics.** For each fit we compute correlation between recovered and true values, separately by parameter type. For matrix-valued parameters ( $A, \xi$ ) we report mean and minimum column correlation; for vector-valued parameters ( $\phi, \psi$ ) we report a single correlation; for person traits  $\theta^{\text{pers}}$  we report mean and minimum per-dimension correlation; for the match shift  $\delta^{\text{mat}}$ , a single vector correlation.

**Results.** Table 3 summarizes recovery across the 20 synthetic datasets. The rater submodel recovers consistently well:  $\phi_r$  correlations exceed 0.95 in

every run,  $\xi_{rj}$  column correlations exceed 0.95 in every run. Item noise  $\psi_j$  recovers moderately, with mean correlation 0.71 at  $D = 3$  and 0.69 at  $D = 4$ .

Loading and person-trait recovery depend on  $D$ . At  $D = 3$ , mean column correlation on  $A$  is 0.84 and mean per-dimension correlation on  $\theta^{\text{pers}}$  is 0.79. At  $D = 4$ , the analogous figures are 0.77 and 0.68, with one run showing a column correlation of  $-0.14$  (a weakly identified column on which MatchAlign could not recover orientation).

The match shift  $\delta^{\text{mat}}$  recovers with mean correlation 0.66 at  $D = 3$  and 0.29 at  $D = 4$ . Four of twenty runs show negative recovery correlations for the match shift, three of these at  $D = 4$ . These are sign-flip artifacts of weakly-identified columns: when MatchAlign cannot uniquely orient one column of  $A$ , the match shift on that column lands at the rotated axis with reversed sign. The sign flips are concentrated in runs where  $A_{\text{corr},\min}$  also shows weak identification, supporting the interpretation that match recovery quality follows column identifiability.

**Chain failure threshold calibration.** The recovery experiments revealed that the chain-failure detection threshold of  $10^{-3}$  originally used during model development was too lax. The synthetic-data setting exposes a deeper form of chain collapse where some chains converge to near-degenerate posteriors with  $s_c$  values around  $10^{-3}$  to  $10^{-2}$ , above the original threshold but clearly distinct from valid chains at  $s_c \gtrsim 0.5$ . Tightening the threshold to  $5 \times 10^{-2}$  excludes these collapsed chains cleanly. This calibration applies to the main analysis as well; in practice it excludes zero or one of eight chains per fit on real data.

**What recovery does and does not validate.** The recovery results justify treating the rater submodel and the loading structure at  $D = 3$  as substantively interpretable. Person-trait recovery is moderate at  $D = 3$  and weaker at  $D = 4$ , consistent with the degradation pattern we observe in PSIS-LOO reliability. The match shift recovers reasonably at  $D = 3$  but unreliably at  $D = 4$ , with sign flips on three of ten runs at  $D = 4$ . We take this as independent support for  $D = 3$  as the primary specification: the substantive quantities the paper relies on, particularly the direction of the match-shift effect, recover reliably at  $D = 3$  and become unreliable at  $D = 4$ .

| Quantity                                     | $D = 3$ | $D = 4$ |
|----------------------------------------------|---------|---------|
| $A$ (column corr., mean)                     | 0.84    | 0.77    |
| $A$ (column corr., min)                      | 0.58    | 0.38    |
| $\theta^{\text{pers}}$ (per-dim corr., mean) | 0.79    | 0.68    |
| $\theta^{\text{pers}}$ (per-dim corr., min)  | 0.70    | 0.56    |
| $\phi$ (rater severity)                      | 0.99    | 0.98    |
| $\xi$ (rater $\times$ item)                  | 0.97    | 0.97    |
| $\psi$ (item noise)                          | 0.71    | 0.69    |
| $\delta^{\text{mat}}$ (match shift)          | 0.66    | 0.29    |

Table 3: Parameter recovery across 10 synthetic datasets per  $D$  value (5 seeds at each of  $P \in \{100, 200\}$ ). Each cell is the mean across runs. The rater submodel recovers consistently; person traits and loadings recover well at  $D = 3$  but degrade at  $D = 4$ . The match-shift correlation degrades sharply at  $D = 4$ , reflecting sign-flip artifacts on runs with weakly identified columns: 1 of 10 runs at  $D = 3$  versus 3 of 10 at  $D = 4$  show negative  $\delta^{\text{mat}}$  correlation, indicating the match shift was recovered onto a rotated axis with reversed sign.

**Chain failure diagnostic.** The MatchAlign procedure assumes each chain has explored a non-degenerate posterior (Xiao et al., 2025). Chains for which the mean posterior std on  $A$  falls below  $5 \times 10^{-2}$  are flagged as collapsed and excluded; in our main fit this excludes zero of eight chains. The synthetic-data setting exposes a deeper form of chain collapse where some chains converge to near-degenerate posteriors with  $s_c$  values around  $10^{-3}$  to  $10^{-2}$ , above the original threshold but clearly distinct from valid chains at  $s_c \gtrsim 0.5$ . Tightening the threshold to  $5 \times 10^{-2}$  excludes these collapsed chains cleanly.

## G Reliability Robustness: High-Agreement Item Subset

To test how far the recovered structure and the domain-match effects depend on the sub-threshold items, we refit the model on the subset of items with Krippendorff’s  $\alpha > 0.5$ : items 1 (3.3.1, decision criteria), 2 (3.1, citation sources), 3 (3.2.1, position statement), 4 (2.3.1, multi-task prompting), and 5 (2.1.2, full task description as prompt). We use  $\alpha > 0.5$  rather than the conventional .667 because the latter retains only three items (1, 3, and the gating-adjacent high- $\alpha$  items), too few to support a three-dimensional latent space;  $\alpha > 0.5$  is the tightest floor that leaves the model estimable. This subset drops the entire active-prompting block (items 7–9) along with the revision item (6), all of which are below threshold.

**Fit and convergence.** We fit at  $D = 3$  with the main sampler settings. All eight chains pass the chain-failure diagnostic, pairwise correlations of chain-mean  $A$  equal 1.000 to three decimals across all 28 pairs after MatchAlign (pivot condition number 2.95), and all rotation-invariant and aligned quantities satisfy  $\hat{R} \leq 1.05$  with  $n_{\text{eff}} \geq 400$ . As in the main analysis, PSIS-LOO is unreliable on this subset (594 observations with  $k > 0.7$ ), so we read the fit structurally rather than for predictive comparison.

**Loading structure.** Table 4 reports the aligned loadings. The delegation block (items 1, 3, 4) reproduces as a single dimension with dominance fractions 0.98, 0.92, and 0.79; the citation item (2) again separates onto its own dimension (dominance 0.83); and item 5 (full task description as prompt) splits into a one-item dimension (dominance 0.94) rather than joining delegation, exactly the behaviour it showed at  $D = 4$  in the main fit (Appendix D). The active-prompting dimension is absent because its only anchor items are the dropped sub-threshold ones; this subset therefore *cannot* evaluate that dimension, and its absence here is structural rather than evidence against it.

**Domain-match effects.** The in-domain citation effect reproduces cleanly:  $+0.35$  ( $P(+) = 1.00$ ) here against  $+0.28$  ( $P(+) = 0.99$ ) in the main fit. The delegation effect attenuates toward null ( $-0.05$ ,  $P(+) = 0.29$  versus  $-0.13$ ,  $P(+) = 0.08$ ); we do not read this as a quantitative replication, because the latent geometry differs and match-shifts are defined relative to a geometry. The strong negative shift the main fit attributed to delegation appears, on this subset, to localise to copy-paste behaviour specifically: item 5’s standalone dimension carries  $-0.32$  ( $P(+) = 0.002$ ), indicating in-domain students reproduce the full task prompt less, while the decision-criteria/position-statement core moves to roughly zero. We flag this as exploratory, since it reads a reconfigured one-item dimension. The prompting match effect ( $+0.29$  in the main fit) is unassessable here.

**What this establishes.** The check partitions the paper’s claims by how much they rest on sub-threshold items. The delegation–citation separation and the in-domain citation effect do not depend on the low- $\alpha$  items: both survive dropping four of nine items. The active-prompting dimension, by contrast, is anchored entirely by sub-threshold

items and rests on the  $\psi_j$  argument of Section 7 alone; this robustness check is structurally unable to corroborate it. The rater submodel behaves consistently with the main fit (rater-severity spread 0.48,  $\sigma_\xi = 0.53$ , 9 of 30 interaction cells credibly nonzero).

| Item      | Cit.        | Deleg.      | Copy        |
|-----------|-------------|-------------|-------------|
| 1 (3.3.1) | 0.74        | <b>4.85</b> | 0.54        |
| 2 (3.1)   | <b>2.56</b> | 1.58        | 0.65        |
| 3 (3.2.1) | −0.23       | <b>3.02</b> | 1.29        |
| 4 (2.3.1) | 1.11        | <b>1.56</b> | 0.52        |
| 5 (2.1.2) | 0.44        | 1.08        | <b>3.35</b> |
| Norm      | 2.93        | 6.23        | 3.72        |

Table 4: Aligned posterior-mean loadings on the  $\alpha > 0.5$  subset at  $D = 3$ . Bold marks each item’s dominant dimension. Delegation (items 1, 3, 4) and citation (item 2) reproduce the main-fit structure; item 5 separates into its own dimension, as at  $D = 4$  in the full fit. The active-prompting dimension is absent because items 7–9 are dropped.

## H Detailed Data and Rating Protocol

### H.1 Study Design

A factorial design crossed expertise level (expert versus novice) with reasoning context (practical versus transdisciplinary) in scenario-based tasks designed around three facets of Critical Online Reasoning: Online Information Acquisition, Reasoning with Evidence, and Critical Information Evaluation (Zlatkin-Troitschanskaia et al., 2023). Participants whose declared major fell within the eight task topics completed one in-domain task (the expert condition;  $m_d = 1$ ) and one out-of-domain task ( $m_d = 0$ ); for example, a medical-major participant performed a medicine-domain task followed by an economics-domain task. We additionally included a control group of students whose major fell outside the eight task topics; these participants completed only out-of-domain tasks ( $m_d = 0$ ) and contribute to item parameters, rater effects, and the population distribution of  $\theta^{\text{pers}}$  but not to the match contrast. Each task was time-boxed to 30 minutes, after which participants were directed to a reflection survey on interest, motivation, and engagement.

Prior to the tasks, AI literacy was assessed using the instrument of Köhler and Hartig (2024) to allow control for individual differences in AI knowledge, use frequency, value, and attitude. All activities were conducted through an anonymized account

on a dedicated virtual machine, ensuring a uniform technical environment across participants. The full protocol, including the AI-literacy pre-test, two tasks, and post-task surveys, took approximately two hours per participant. Participants received monetary compensation. Tasks were drawn from eight academic topics balanced across the medicine and economics domains.

## H.2 Annotation Schema

Coders annotated each task interaction using a schema with three dimensions: ChatGPT Use (a single gating item), Prompting Behavior (organized into three sub-dimensions: Prompt Content, Prompting Strategies, Prompt Granularity), and ChatGPT Contribution to the Final Output (three ordinal items, one per response field). The schema was applied separately to the practical and transdisciplinary reasoning contexts of each task, so each participant’s interaction log produced two coded units. Item codes follow the rubric hierarchy: e.g., 2.1.2 is the second sub-dimension of dimension 2.

The 15 items in the original schema split into binary and four-point ordinal forms. The binary items in the Prompt Content sub-dimension (2.1.1–2.1.5) are ordered conceptually by increasing cognitive engagement: from copying the task description verbatim into the prompt at the low end, through asking concrete self-formulated questions, to asking abstract conceptual questions at the high end. The Prompting Strategies sub-dimension (2.2.1–2.2.4) captures more deliberate interaction patterns including correction, question-back, source-grounded text generation, and revision requests. The Prompt Granularity items (2.3.1, 2.3.2) record whether multiple distinct prompts were issued for different tasks or for different aspects of a single task. The three contribution items (3.1, 3.2.1, 3.3.1) are four-point ordinal scales recording the degree to which ChatGPT-generated content was incorporated into the source citation field, the position statement, and the decision criteria, ranging from no use through verified use to wholesale reproduction.

## H.3 Coding Procedure

A calibration round preceded the main coding phase, in which two coders independently annotated a subset spanning two domains, four tasks, five documents, and two reasoning contexts. The calibration yielded an average Cohen’s  $\kappa$  of .74 in the medical domain and .68 in the economics

domain (Cohen, 1960). Items with low calibration agreement were discussed among all coders using concrete examples until shared interpretations were reached.

Two coders were assigned per domain, with practical and transdisciplinary reasoning contexts distributed between them within the economics domain. Apart from the economics practical reasoning context, where one coder was unable to complete the assignment, all documents were coded by at least two independent coders. The full corpus comprised 958 documents. Items 2.1.4, 2.1.5, and 3.1 were further adjudicated by a third coder for documents on which the first two coders had disagreed (263 documents in total: 111 economics, 152 medical), with the third coder’s rating used as a tiebreaker for the binary items (majority) and entering the rounded mean for the ordinal item 3.1.

## H.4 Item Set for the Present Analysis

We exclude four items from analysis on the basis of insufficient observation count or near-zero base rates: 2.2.1 (asking ChatGPT to correct one’s own statements) and 2.2.2 (instructing ChatGPT to pose questions) were excluded due to low observation count; 2.2.3 (instructing ChatGPT to formulate text on the basis of provided sources) showed near-zero base rates and the lowest  $\alpha$  in the schema; 2.1.3 (copying a partial task excerpt) was excluded due to high conceptual overlap with 2.1.2 combined with low base rates and low  $\alpha$ . We also exclude the gating item 1.1 itself because the analysis is restricted to documents in which ChatGPT was consulted, where 1.1 is constant by construction, and 2.1.1 because it took the value 1 in nearly all retained documents and carries negligible variance.

The remaining nine items form the analysis set: the polytomous contribution items 3.3.1, 3.1, and 3.2.1, the binary granularity items 2.3.1 and 2.3.2, and the binary content and strategy items 2.1.2, 2.1.4, 2.1.5, and 2.2.4. Of these, six (3.1, 2.3.1, 2.1.4, 2.1.5, 2.2.4, and 2.3.2) have  $\alpha$  below the .667 threshold for tentative conclusions. We retain them on three grounds. First, the model includes an item-level noise parameter  $\psi_j$  that scales the response function for each item independently; items with greater rater disagreement are accommodated through larger  $\psi_j$  rather than treated as fully reliable signal, so the model down-weights their contribution to the latent traits accordingly. Second, the rater-by-item interaction  $\xi_{rj}$  absorbs structured rater disagreement, separating system-

atic differences in how individual raters apply each item from random noise; Krippendorff's  $\alpha$  does not make this distinction and treats both as disagreement. Third, in the fitted posterior, the  $\psi_j$  values for these six items (3.1:  $\psi = 1.03$ ; 2.3.1:  $\psi = 0.91$ ; 2.1.4:  $\psi = 0.77$ ; 2.1.5:  $\psi = 0.93$ ; 2.2.4:  $\psi = 1.02$ ; 2.3.2:  $\psi = 0.93$ ) all lie at or below 1.05, close to the cross-item mean of 0.95, suggesting the model does not flag them as noisier than the higher- $\alpha$  items once rater effects are partialled out. We interpret this as evidence that much of what  $\alpha$  records as disagreement on these items reflects systematic rater style rather than random measurement error, and is therefore recoverable in a model that explicitly represents raters.

## I Item-Level Results

### I.1 Loading Matrix at $D = 3$

Table 5 reports the posterior mean and standard deviation of every entry of the aligned loading matrix  $A$  at  $D = 3$ . Bold entries mark the dominant dimension for each item. Column norms in the bottom row order the dimensions by total loading magnitude:  $D_2$  (5.97) accounts for the largest share of variance, followed by  $D_1$  (2.99) and  $D_3$  (2.77).

### I.2 Item-Level Noise

Table 6 reports the posterior mean and 95% credible interval of the item noise parameter  $\psi_j$ . Item 1 (3.3.1, decision criteria) has the highest noise at  $\psi = 1.16$ , slightly above the cross-item mean of 0.95. The six items with sub-threshold Krippendorff's  $\alpha$  (3.1, 2.3.1, 2.1.4, 2.1.5, 2.2.4, 2.3.2) all have  $\psi$  within 0.2 of the cross-item mean of 0.95, the largest deviation being item 2.1.4 at 0.77.

### I.3 Rater Severity

Rater severities at  $D = 3$  have mean values  $\phi = (-0.39, +0.42, +0.19, +0.22, -0.15, -0.29)$ , a cross-rater spread of  $\hat{\sigma} = 0.29$ , and a range of 0.80 logits from the most lenient to the harshest rater. Differences of this magnitude motivate the explicit rater submodel.

### I.4 Rater $\times$ Item Interaction

Figure 4 visualizes the full rater-by-item interaction  $\xi_{rj}$  at  $D = 3$ , with the posterior mean (left), signal-to-noise ratio (centre), and rater main effects  $\phi_r$  for context (right). 27 of 54 cells cross the  $|\xi|/\text{sd}(\xi) > 2$  threshold for credibly nonzero. The maximum effect is  $|\xi_{2,4}| = 1.52$ .

## I.5 Domain Match

Figure 5 shows the posterior of the domain match effect  $\delta_{\text{mat}}$  broken down by latent dimension. Among students whose major is represented in the task topics, working in one's own academic domain shifts students toward higher values on  $D_1$  (+0.29,  $P(+) = 1.00$ ) and  $D_3$  (+0.28,  $P(+) = 0.99$ ); the effect on  $D_2$  is credibly negative ( $-0.13$ ,  $P(+) = 0.08$ ).

| Item      | $D_1$              | $D_2$              | $D_3$              |
|-----------|--------------------|--------------------|--------------------|
| 1 (3.3.1) | -0.04 (0.11)       | <b>4.22</b> (0.83) | 0.61 (0.43)        |
| 2 (3.1)   | -0.05 (0.08)       | 1.62 (0.38)        | <b>2.36</b> (0.56) |
| 3 (3.2.1) | 0.47 (0.14)        | <b>2.95</b> (0.49) | -0.29 (0.39)       |
| 4 (2.3.1) | 0.55 (0.12)        | <b>1.89</b> (0.40) | 1.23 (0.33)        |
| 5 (2.1.2) | -0.36 (0.10)       | <b>0.95</b> (0.18) | 0.14 (0.16)        |
| 6 (2.2.4) | 0.77 (0.18)        | <b>1.26</b> (0.24) | -0.25 (0.19)       |
| 7 (2.1.4) | <b>1.35</b> (0.23) | 0.30 (0.10)        | 0.24 (0.08)        |
| 8 (2.1.5) | <b>1.76</b> (0.33) | -0.07 (0.13)       | 0.02 (0.10)        |
| 9 (2.3.2) | <b>1.66</b> (0.30) | 0.57 (0.14)        | -0.08 (0.10)       |
| Norm      | 2.99               | 5.97               | 2.77               |

Table 5: Posterior mean and standard deviation of the aligned loading matrix  $A$  at  $D = 3$ . Column norms in the bottom row indicate per-dimension variance contribution. Bold entries mark the dominant dimension for each item. Item codes in parentheses correspond to the rubric identifiers in Appendix H.

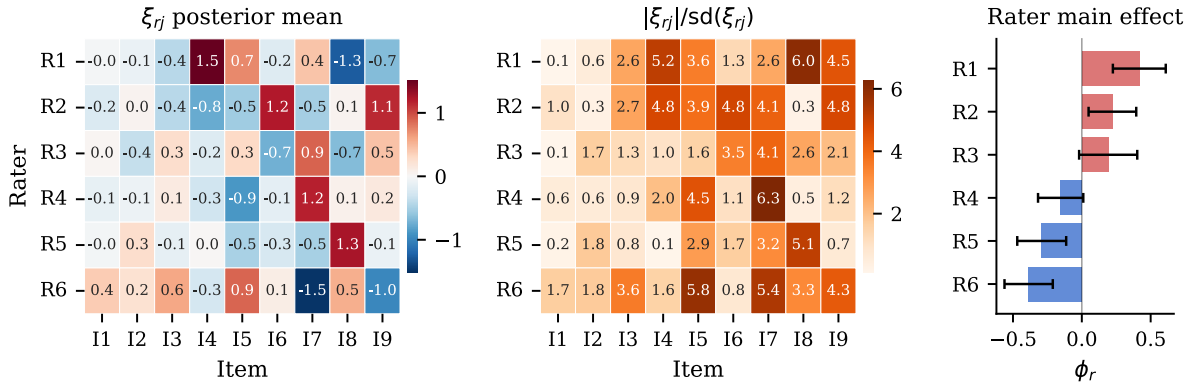


Figure 4: Rater  $\times$  Item interaction  $\xi_{rj}$  at  $D = 3$ . Left: posterior mean. Centre: signal-to-noise ratio  $|\xi_{rj}|/\text{sd}(\xi_{rj})$ , with cells exceeding 2 credibly nonzero. Right: rater main effects  $\phi_r$  for comparison.

| Item      | $\psi_j$ | 95% CI       |
|-----------|----------|--------------|
| 1 (3.3.1) | 1.16     | [0.83, 1.57] |
| 2 (3.1)   | 1.03     | [0.73, 1.40] |
| 3 (3.2.1) | 0.80     | [0.59, 1.05] |
| 4 (2.3.1) | 0.91     | [0.65, 1.22] |
| 5 (2.1.2) | 1.01     | [0.71, 1.37] |
| 6 (2.2.4) | 1.02     | [0.72, 1.39] |
| 7 (2.1.4) | 0.77     | [0.56, 1.03] |
| 8 (2.1.5) | 0.93     | [0.66, 1.24] |
| 9 (2.3.2) | 0.93     | [0.66, 1.24] |

Table 6: Item-level rater noise  $\psi_j$  at  $D = 3$  with 95% posterior credible intervals. No item exceeds  $\psi > 1.2$ .

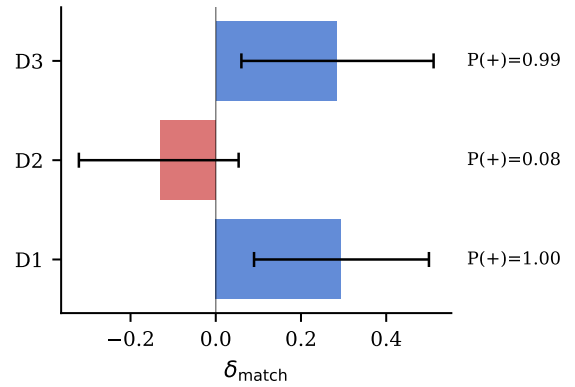


Figure 5: Domain match effect  $\delta_{\text{mat}}$  per latent dimension at  $D = 3$ . Posterior means with 95% credible intervals;  $P(+)$  annotations indicate posterior probability of a positive effect.