Hierarchical Bayesian Multi-Dimensional IRT Applied to 200k Concept Tests

Martin Segado *, Aaron Adair *, Atharva Dange *, Miao Yi Deng, David Pritchard *

Posted Date: 14 May 2026

doi: 10.20944/preprints202605.0952.v1

Keywords: hierarchical bayesian modeling; variational inference; multi-dimensional analysis; multiple choice tests; physics education research; item response theory; data cleaning; psychometrics

 doi:10.20944/preprints202605.0952.v1

Article

Hierarchical Bayesian Multi-Dimensional IRT Applied to 200k Concept Tests

Martin Segado *, Aaron Adair *, Atharva Dange *, Miao Yi Deng and David Pritchard *

Massachusetts Institute of Technology, Cambridge, MA 02139, USA
* Correspondence: dpritch@mit.edu (D.P.), dange98@mit.edu (A.D.), adairaar@mit.edu (A.A.), msegado@mit.edu (M.S.)

Abstract

We report the use of our group's hierarchical Bayesian implementation of the Multi-dimensional Nominal Categories Model followed by standard factor rotations of the principal dimensions to obtain 29 curated sparse dimensions from a set of 203,564 (104,998 pre and 98,566 post) administrations of a multiple-choice concept test in mechanics. We emphasize our careful attention to issues common to fitting such multi-parameter models to large data sets: a novel set of filters to remove administrations from non-conscientious testees, use of Bayesian methods to avoid overfitting, selecting the best transformations to find easily identifiable sparse dimensions, and verification and pruning of these using bootstrap samples. We demonstrate that most dimensions are invariant across different demographically different samples of students as well as between pre-instruction vs post-instruction samples. Most sparse dimensions correspond to well-known misconceptions in mechanics.

Keywords: hierarchical bayesian modeling; variational inference; multi-dimensional analysis; multiple choice tests; physics education research; item response theory; data cleaning; psychometrics

---

1. Introduction

The rapid growth of data collection in modern scientific and technological applications has intensified interest in Bayesian methods for high-dimensional statistical problems. In many domains—including genomics, neuroscience, machine learning, and the social sciences—researchers confront large datasets with complex dependence structures and latent variables whose dimensionality may itself be uncertain. Bayesian inference offers a flexible framework for modeling such complexity, facilitating principled uncertainty quantification, incorporation of prior information, and model comparison. However, these advantages come with substantial computational and methodological challenges: selecting appropriate priors, achieving scalable posterior computation, and ensuring reliable inference in high-dimensional parameter spaces.

Large-scale multiple-choice assessments in educational measurement and psychometrics routinely collect responses from thousands of testees on dozens of multiple-choice items, yielding high-dimensional categorical data. Such datasets are typically analyzed using one-dimensional item response theory (IRT) models to estimate latent overall ability, generally collapsing responses into dichotomous correct/incorrect outcomes and thus discarding information contained in the particular distractor choices ([1]). In research-designed multiple-choice instruments, testees exhibiting persistent procedural errors or other systematic response patterns may select characteristic combinations of distractors. Therefore high-dimensional Bayesian modeling of all response categories offers the possibility of revealing underlying latent dimensions. In certain concept tests, such as the Force Concept Inventory (FCI), robust patterns of these latent dimensions have been shown to reveal misconceptions known from previous investigations ([2]).

Here, we demonstrate an application of a hierarchical Bayesian implementation of the Multi-dimensional Nominal Categories Model (MNCM) from IRT to infer latent dimensions underlying high-dimensional categorical response data. In this implementation, item-category parameters share

 doi:10.20944/preprints202605.0952.v1

shrinkage priors governed by higher-level hyperparameters, and examinees' latent misconception profiles are modeled as multivariate abilities. This hierarchical structure improves estimation in high-dimensional settings by sharing information across items and enables automatic selection of the effective latent dimensionality. To make inference feasible for large datasets, we used a variational inference approach that approximates the joint posterior distribution while remaining computationally tractable. Applying the proposed approach to a dataset of 203,564 administrations of the FCI (∼181k after cleaning), we identify a 26-dimensional latent structure.

Interpreting this many dimensions is a challenging problem, as responses often exhibit complex dependency structures. To address this issue, we identify transformations of the latent space that yield "sparse dimensions", in which each dimension loads heavily on only a few distractors. We show that sparse representations substantially improve interpretability of the inferred latent structure in applications such as concept inventories by identifying most of these sparse dimensions as intellectually coherent misconceptions. More generally, such rotations reframe the problem of interpreting the output of a complex multi-dimensional statistical model as the problem of identifying interpretable sparse structures in high-dimensional categorical response data.

An overview of our full analytical pipeline is shown in Figure 1. It illustrates the process by which raw response vectors proceed through cleaning, hierarchical Bayesian estimation, sparse rotation, and universality validation, ultimately leading to the identification of interpretable misconception dimensions. These developments are detailed in the following sections: in Section 2, we describe our collection of ∼204k responses from a concept test followed by a simple method to filter out responses from non-conscientious testees; in Section 3, the hierarchical Bayesian model for the Multi-dimensional Nominal Categories Model and its variational estimation is presented; and in Section 4, we present the rotation-based sparse representation, its validation across bootstrap and independent subsets, and demonstrate that these sparse vectors provide reliable inference in high-dimensional categorical data by connecting most of them with known misconceptions. In doing so, we contribute to the broader aims of this Special Issue by showcasing a valuable application of Bayesian hierarchical modeling and scalable inference that provides meaningful interpretation for high-dimensional analysis of educational data.

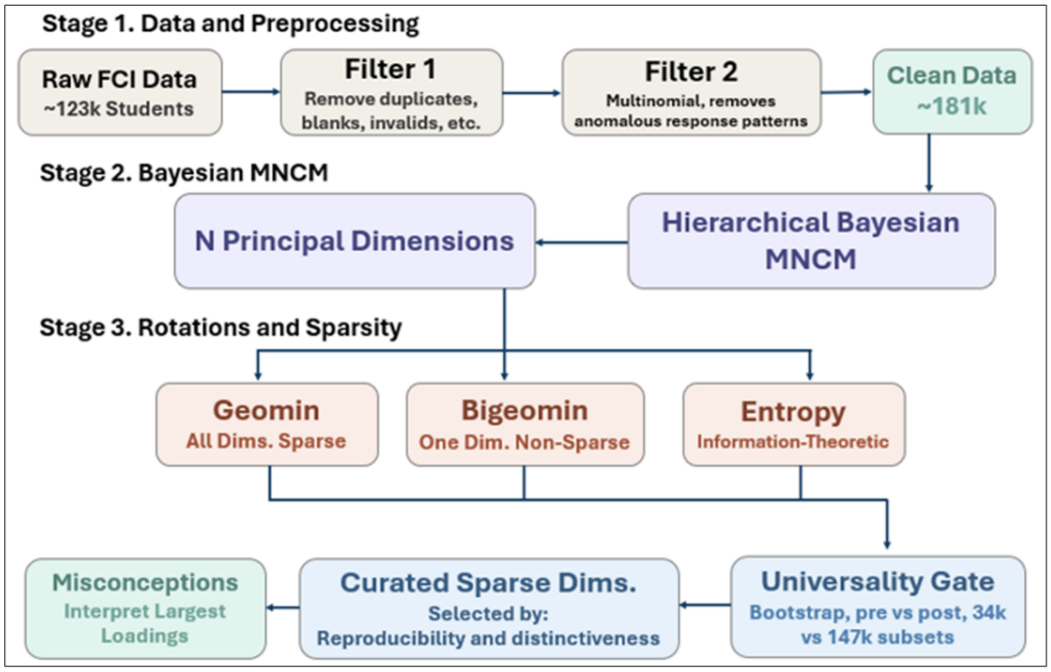

Figure 1. Overview of the analysis pipeline, from raw FCI data through filtering, hierarchical Bayesian MNCM estimation, sparse rotation, and universality validation, to identified misconceptions.

 doi:10.20944/preprints202605.0952.v1

2. Data and Preprocessing

We analyze data from the 1995 revision of the Force Concept Inventory ([3,4]), a pioneering concept test designed to probe student understanding of the fundamental principles of Newtonian mechanics. The FCI consists of 30 multiple-choice questions, each with five response choices (categories) including one correct answer and several distractors intentionally designed to reflect common student misconceptions about force and motion. Results obtained with this instrument showed that most students finishing an introductory Newtonian mechanics course lacked the ability to apply Newton's laws to everyday circumstances and that standard lecture-based instruction improved their ability to do this by less than 25% ([5]). Because of its diagnostic value, the FCI has become a central instrument in Physics Education Research (PER) and has been used extensively to study instructional effectiveness, conceptual change, and persistent misconceptions in introductory mechanics courses ([5–7]). Student responses have therefore been collected by discipline-based organizations and research groups as part of efforts to measure and improve instruction. These data sources are described in Section 2.1 below, and the steps that we have taken to clean the data are described in Section 2.2

2.1. Data Set Description and Origins

By combining test results from numerous sources, several of whom collected data from classrooms across a wide variety of schools, we have obtained data from 203,564 administrations of the FCI, the vast majority from pre- and post-instruction administrations by the teacher in charge of a particular college course. A smaller fraction originates from high-school implementations of the FCI and workshops involving pre-service or in-service teachers. The data from each administration constitute a response vector of length 30, where each component is A to E or some other symbol to indicate "no response" to that question.

2.2. Cleaning the Data: Identification and Removal of Problematic Response Vectors

Most large data sets, especially those collected from multiple teachers at multiple schools, must be checked for errors and noise: duplicates, deliberate cheating by the subjects or others along the chain of collection, student use of heuristic test-taking strategies, signs of students running out of time, and responses by students who lacked the motivation or the honesty to enter data conscientiously. Most of these sources of error result in the student emphasizing or de-emphasizing some categorical responses or leaving blank responses.

To obtain a clean dataset suitable for analysis, we applied two Filters. With Filter 1, we removed response vectors that are clearly duplicates, those containing any invalid responses (entries other than the valid answer choices A–E), and those with missing components (often indicated by *, "NA", or "0"). These records were removed so that each remaining administration consisted only of valid responses to all 30 items. After applying Filter 1, the data had 191k usable FCI administrations (97,453 pre-instruction and 93,390 post-instruction tests), with the results in Table 1.

 doi:10.20944/preprints202605.0952.v1

Table 1. Summary of the FCI datasets used in this study and the sequential preprocessing steps applied to them. All counts are numbers of unique students. Raw Students lists the number of student records initially available from each dataset. Filter 1 retains administrations with complete pre- and post-test responses. Filter 2 applies a likelihood-based screen that removes administrations with anomalous distributions of answer choices across the 30 FCI items. The final columns report the mean pre-test score, mean post-test score, and normalized gain for each paired dataset. For further details on data provenance and access, see the Data Availability Statement.

Dataset	Subset	Raw	Filter 1		Multinomial		Mult. Removed (%)			Pre	Post	Norm.
			Pre	Post	Pre	Post	Pre	Post	Comb.			Gain
PhysPort	More Selective	22,551	12,452	17,864	11,866	10,942	4.7	38.7	24.8	13.80	18.49	0.29
	Selective	29,548	24,628	20,727	23,801	19,807	3.4	4.4	3.9	11.56	17.20	0.31
	Inclusive	12,034	8,861	6,934	8,648	6,923	2.4	0.2	1.4	12.88	17.14	0.25
	2yr & Other	4,629	3,755	3,299	3,722	3,279	0.9	0.6	0.8	13.57	17.39	0.23
	US High School	6,935	5,467	4,764	5,452	4,753	0.3	0.2	0.3	9.19	17.06	0.38
	International	13,314	10,832	8,592	10,785	8,574	0.4	0.2	0.3	12.91	17.40	0.26
	Subtotal	**89,011**	**65,995**	**62,180**	**64,274**	**54,278**	**2.6**	**12.7**	**7.5**	**12.28**	**17.47**	**0.28**
LASSO		15,751	14,777	14,550	14,646	14,285	0.9	1.8	1.4	11.39	17.23	0.28
Modeling		10,039	8,941	8,941	8,932	8,940	0.1	0.0	0.1	8.09	16.32	0.39
Central Southwest		4,363	4,359	4,359	4,350	4,358	0.2	0.0	0.1	12.29	21.43	0.52
Western Community College		195	193	190	193	190	0.0	0.0	0.0	13.55	20.55	0.47
Eastern Selective		165	71	71	71	71	0.0	0.0	0.0	20.77	26.66	0.59
AMTA		164	164	164	163	162	0.6	1.2	0.9	22.74	24.95	0.26
Eastern Private		159	150	132	150	132	0.0	0.0	0.0	8.23	12.39	0.19
Misc		3,260	2,803	2,803	2,801	2,790	0.1	0.5	0.3	11.30	16.13	0.27
TOTAL		**123,107**	**97,453**	**93,390**	**95,580**	**85,206**	**1.9**	**8.8**	**5.3**	**11.61**	**17.49**	**0.31**

Filter 2 is designed to eliminate response vectors that have excesses or deficits of particular responses that often result from heuristic test-taking strategies (e.g., always answer A or B when in doubt) or low-effort responding. This filter is based on a five-dimensional "fingerprint" of each response vector that consists of the number of responses of each category A, B, C, D, and E. We observe that in the majority of cases, these counts are well-described by a simple multinomial distribution (see Figure 2). We therefore apply a multinomial test to identify significant outliers. To avoid the outliers influencing the parameters of the proposed distribution, we iteratively discard outliers and re-fit the distribution on only non-outlier points, repeating this process until the results converge on a consistent set of retained data.

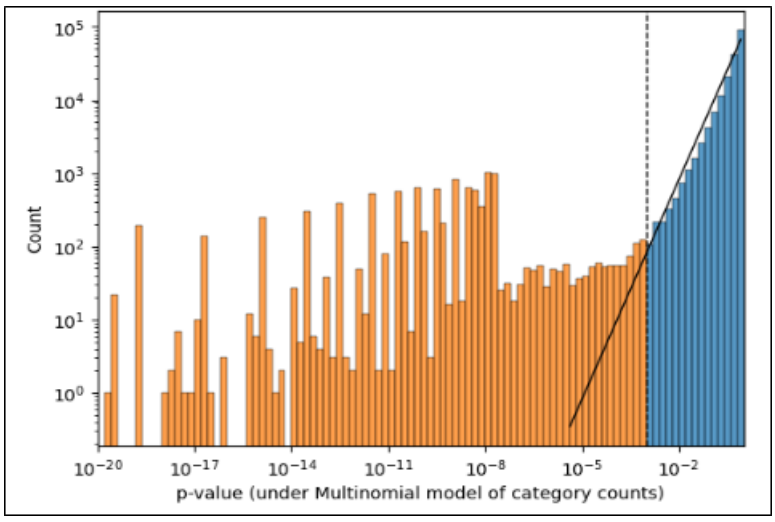

Figure 2. Distribution of response fingerprint p-values under a multinomial test. The solid line shows the expected distribution under the null hypothesis. Response vectors with p-values smaller than 0.001 were eliminated. For better visibility of the retained bins, values smaller than 10^{-20} are not shown.

 doi:10.20944/preprints202605.0952.v1

Overall, Filter 2 (non-conscientious behavior) resulted in the elimination of $\sim 1\%$ of the data in all of the data sets in Table 1, but for the PhysPort data, it was $\sim 7.5\%$ Clearly, the More Selective subset of the PhysPort data exhibits a disproportionately high exclusion rate relative to all other subsets, and its excluded attempts cluster densely near vertex A, or in mixed clusters of As and Bs, indicating systematic over-selection of that choice. In Figure 3, the dotted line between "all A" and "all B" vertices corresponds to students toggling randomly between A and B for every response and accounts for the response fingerprints with over 100 instances and probability below 10^{-8}. Other excluded attempts follow the broader pattern of over-selection near vertices A and C seen across the full dataset. Whether students at more selective institutions are systematically less motivated to engage conscientiously with the FCI, and the underlying reasons for this, remains an open question for future study.

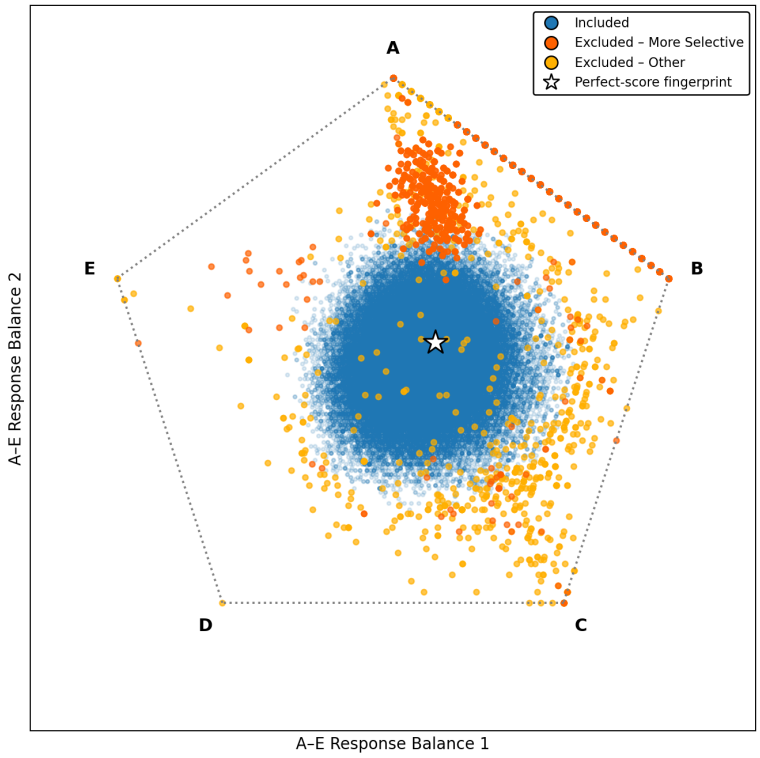

Figure 3. Pentagon projection of response-vector balance scores for all test attempts. Each point represents one attempt mapped onto a two-dimensional pentagon basis, where vertices A–E correspond to the five answer choices. Blue points are retained attempts; dark orange points are attempts excluded from the More Selective institutional subset, and yellow points are excluded attempts from all other groups. The white star marker indicates the perfect-score fingerprint.

3. Bayesian Model Formulation and Inference

3.1. The Multidimensional Nominal Categories Model

The Multidimensional Nominal Categories Model (MNCM; [8], [9]) generalizes Bock's one-dimensional nominal categories model to D latent dimensions ([10]). The ability of student s on dimension d is denoted by $\theta_s^{(d)}$, and each dimension d loads on each item-category pair with slope $a_d^{(i,c)}$ to give a category "tendency",

$$t_s^{(i,c)} = \sum_d a_d^{(i,c)} \theta_s^{(d)} + b^{(i,c)}, \tag{1}$$

 doi:10.20944/preprints202605.0952.v1

where each item-category pair also has an intercept parameter $b^{(i,c)}$ which allows the baseline response probabilities to differ across categories. The predicted probability that student s selects category c for item i is then given by the multinomial logistic function,

$$p\left(r^{(i)} = c \mid \mathbf{t}_s^{(i)}\right) = \frac{\exp\left(t_s^{(i,c)}\right)}{\sum_{c'} \exp\left(t_s^{(i,c')}\right)}. \tag{2}$$

This model is used in several fields and is also known as the Multidimensional Nominal Response Model (MNRM) ([11]), Nominal Factor Analysis ([12]), the Multidimensional Generalization of the Nominal Categories Model ([13]), and MAXMC ([14]). The formulation is mathematically equivalent to a multinomial logistic regression with D latent predictors.

Combining the response tendencies across all items and categories into matrix notation allows for a more compact representation of the model parameters. Let

- $\boldsymbol{T} \in \mathbb{R}^{S \times IC}$ denote the matrix of student response tendencies, where each row corresponds to a student s and each column to an item's category (i, c)
- $\boldsymbol{\Theta} \in \mathbb{R}^{S \times D}$ denote the matrix of student latent ability vectors, where each row corresponds to a student s and each column to a latent dimension d
- $\mathbf{A} \in \mathbb{R}^{D \times IC}$ denote the matrix of category slope vectors across all items and categories
- $\mathbf{b} \in \mathbb{R}^{1 \times IC}$ denote the vector of category intercepts

(Note that in this notation, dimensions are associated with *rows* of $\mathbf{A}$ rather than columns.)

The MNCM is subject to multiple classes of indeterminacy: location invariance of the abilities, location invariance of the item parameters, scale invariance, and rotation invariance. These invariances take some additional consideration to address compared to a fully-identified model, but also allow for transformations which yield more interpretable results.

In the first case, shifting all student abilities on any dimension d leaves the predicted response probabilities unchanged, provided intercepts are adjusted accordingly. We resolved this indeterminacy by centering $\boldsymbol{\Theta}$ at zero across the student sample for every dimension, compensating with the following adjustment to the intercepts:

$$\Delta b^{(i,c)} = \sum_d a_d^{(i,c)} \left\langle \theta^{(d)} \right\rangle_s. \tag{3}$$

where $\langle \cdot \rangle_s$ represents the arithmetic mean over students.

In the second case, an examination of the model formulation will reveal that the slope and intercepts parameters for each item are partly redundant: all may be shifted (independently per item, and per dimension for the slopes) by the same additive constant without changing the probabilities. To eliminate these shift-invariances, two traditional approaches exist: simple constraints (fixing the slope and intercept of a clear reference category to zero where one exists, such as for the keyed correct-answer category) or deviation constraints (centering category parameters within each item such that $\langle a_d^{(i,c)} \rangle_c = 0$ and $\langle b^{(i,c)} \rangle_c = 0$). We apply simple constraints for ease of interpretation, yielding slope and intercept parameters for the distractors whose values may be understood relative to the correct answer in each item.

Finally, any invertible linear transformation applied to $\boldsymbol{\Theta}$ (with the corresponding inverse transformation applied to $\mathbf{A}$) also preserves the predicted response probabilities. This combined scale-and-rotation freedom is initially resolved by constraining the sample covariance matrices of $\boldsymbol{\Theta}$ and $\mathbf{A}$ to be identity and diagonal, respectively, and by ranking dimensions in order of decreasing slope variance.

The indeterminacies discussed above are only softly-constrained while fitting the model (by our choice of Bayesian prior, discussed below), with the exact identification constraints imposed analytically as a post-processing step.

 doi:10.20944/preprints202605.0952.v1

3.2. Bayesian Hierarchical Modeling

We embed MNCM in a hierarchical Bayesian framework as suggested by [15]. Global scale hyperparameters α_d and β govern the prior standard deviations of item slopes and intercepts, respectively, allowing the data to determine the effective strength of each latent dimension. The full generative model (4)–(7) uses a weakly informative half-Cauchy hyperprior for these parameters. This distribution has wider tails than a normal distribution—allowing the model to sample widely on α_d and β—while its non-vanishing density near the origin allows the model to shrink the scale of unnecessary dimensions to ∼zero, acting as a soft regularizer on model dimensionality.

$$\alpha_d, \beta \sim \text{HalfCauchy}(5), \tag{4}$$

$$a_d^{(i,c)} \sim \mathcal{N}(0, \alpha_d), \tag{5}$$

$$b^{(i,c)} \sim \mathcal{N}(0, \beta), \tag{6}$$

$$\theta_s^{(d)} \sim \mathcal{N}(0, 1). \tag{7}$$

The prior means of all item parameters are fixed at zero (noting that the invariances in the model make this choice arbitrary), and student abilities $\theta^{(d)}$ are assigned standard normal priors consistent with conventional IRT practice.

Bayesian inference requires solving Bayes' law for the posterior distribution over all latent variables, $\Omega = \{\boldsymbol{\Theta}, \mathbf{A}, \mathbf{b}, \boldsymbol{\alpha}, \beta\}$, given the observed set of responses R:

$$p(\Omega \mid R) = \frac{p(R \mid \Omega)\, p(\Omega)}{p(R)}. \tag{8}$$

Unfortunately, this is analytically intractable. Markov Chain Monte Carlo (MCMC) is a common approximation method, but it is computationally expensive. Instead of MCMC, we adopt Stochastic Variational Inference (SVI), which reformulates posterior inference as an optimization problem. Rather than computing the posterior directly, SVI fits an analytically-tractable approximation to the posterior distribution, $q(\Omega) \approx p(\Omega)$, by maximizing a surrogate objective called the Evidence Lower Bound (ELBO).

Our choice of $q(\Omega)$ follows a mean-field normal approximation operationalized in the following way: each real-valued latent variable is assigned an independent univariate Gaussian posterior characterized by a mean and standard deviation, both of which are optimized during inference. For positive-constrained parameters (α_d, β, and all posterior standard deviations), surrogate unconstrained variables are introduced internally and mapped to positive values via the softplus transformation: $x \mapsto \log(1 + e^x)$.

Optimization proceeds via the Adam optimizer with a learning rate schedule that decreases non-linearly from 0.05 to zero over 30,000 steps. Parameter initialization follows an adaptation of the fast IRT approximation described by Zhang et al. (2020) to improve convergence speed. Final point estimates are obtained as expected a posteriori (EAP) values extracted from the fitted variational distributions, with posterior (co)variances retained to quantify parameter uncertainty.

4. Interpreting Latent Dimensions: Rotations and Misconceptions

A central problem with procedures that extract multiple dimensions or factors (as with exploratory factor analysis) from large data sets is interpreting the results so that they are understandable and informative. In fact, this problem was our motivation for our research agenda: we wanted to discover and study student misconceptions using a large sample of response vectors from the FCI, with the ultimate objective of developing specific remediative pedagogies. These desires motivated our selection of the MNCM model and the development of the methods described in this paper (and [2])—all based on the hopeful verification of our hypothesis that misconceptions would manifest as a small set of wrong answers ("distractors") on the FCI (or any other research-designed concept test) that would

 doi:10.20944/preprints202605.0952.v1

be preferentially selected by students harboring that misconception. Since the questions on the FCI cover numerous misconceptions we expected that a single misconception would "load" on only a few distractors. This is achieved by rotating the principal vectors found by the MNCM model to achieve *sparse dimensions* that load heavily on just a few distractors.

In this section, we first identify a curated set of sparse dimensions from multiple rotated solutions, then demonstrate that they are highly universal, i.e., invariant across various different sub-samples of our data (including those with different levels of instruction or student demographics). We show that most of these can be identified as misconceptions, both novel and previously known from the misconceptions literature.

4.1. Rotations find Sparse Dimensions

The principal dimensions discovered by the hierarchical Bayes procedures described above typically load significantly on many distractors across multiple questions and therefore are not interpretable as a single misconception. The key to finding sparse dimensions is to transform these principal dimensions so that each transformed dimension loads heavily on only a few distractors, thus facilitating the identification of these new *sparse dimensions*. This is allowed because the tendencies (Equation 1) that predict the probabilities of a student's responses are invariant under joint rotation of both the student ability vectors and the loading vectors (θ and a) since the corresponding tendencies are proportional to the *scalar product* of θ and a. This general result, already used in standard factor analysis, is apparent when looking at the tendency equation in matrix form:

$$\boldsymbol{T} = \boldsymbol{\Theta A} + \boldsymbol{b} = \boldsymbol{\Theta}(\boldsymbol{R}\boldsymbol{R}^{-1})\boldsymbol{A} + \boldsymbol{b} = (\boldsymbol{\Theta R})(\boldsymbol{R}^{-1}\boldsymbol{A}) + \boldsymbol{b} \tag{9}$$

Our group's exploration of rotation methods to find sparse vectors ([17]) explored roughly a dozen to find those that produce the most universal dimensions that clearly correspond to one (or two) intellectually consistent misconceptions. Here, we limit ourselves to the most promising transformation method in each of three major categories: an orthogonal method (the minimum-entropy rotation, which has its basis in information theory), an oblique method (geominQ), and an oblique bi-factor method (bigeominQ) which explicitly permits a single non-sparse dimension. Each of these is applied 100 times with random initial rotation matrices in order to fully explore the solution space, which generally contains multiple informative local minima of the rotation criterion. The best rotated solution for each of these methods (designated by a "-0" suffix) are compared with the principal dimensions in Figure 4.

 doi:10.20944/preprints202605.0952.v1

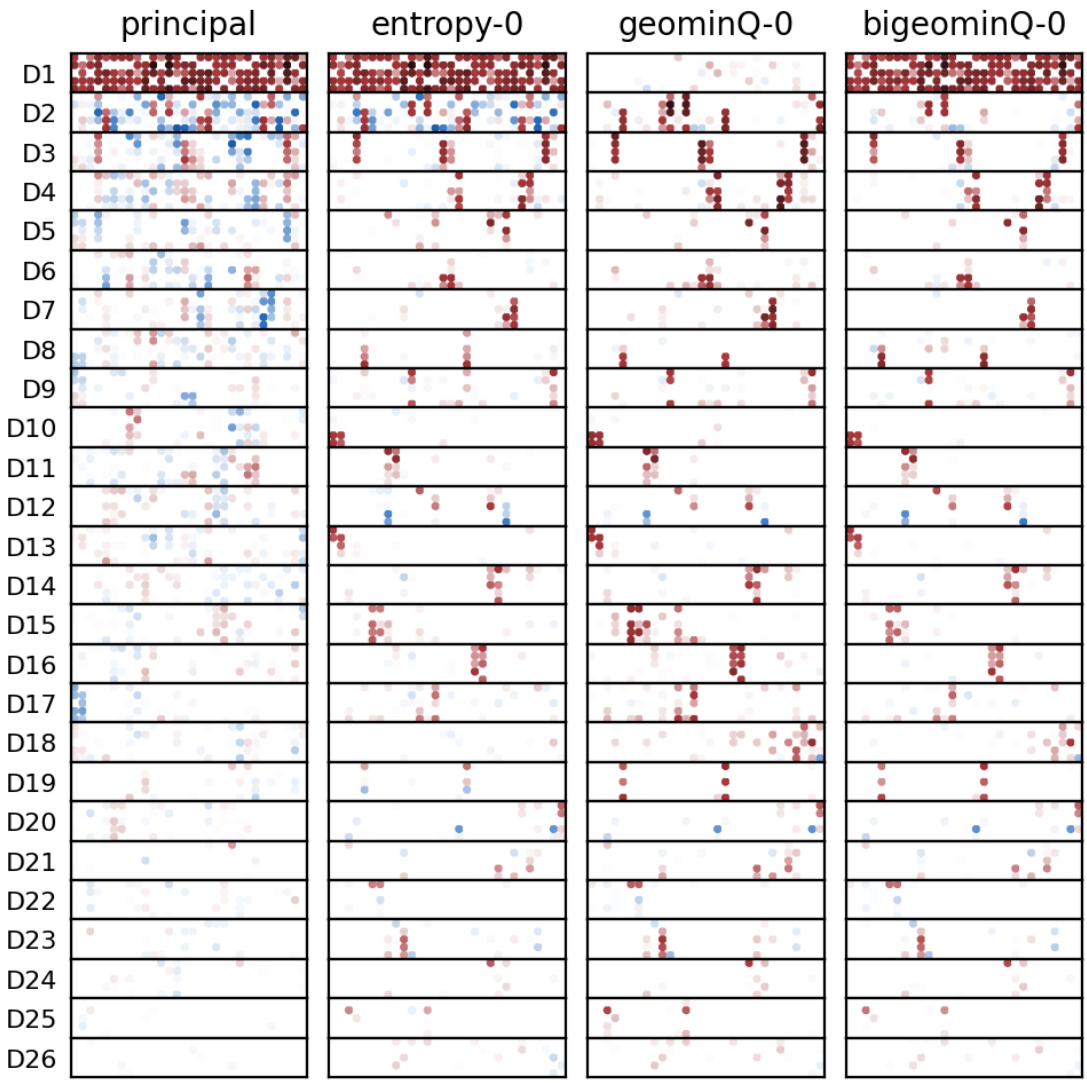

Figure 4. Principal and various Sparse Rotations. Here we show only one of several local optima found for each sparse rotation (the ones achieving the lowest rotation criterion, designated by the suffix "-0"). Due to the rotation invariance of the MNCM, the ordering of the sparse dimensions is somewhat arbitrary—we attempt to match these as closely as possible in our presentation for ease of comparison.

The first principal dimension loads negatively on every distractor, clearly identifying it as "Newtonian Incorrect". Functionally-identical versions of this dimension are also found by two of our three factor rotation methods; its dominance confirms that the FCI is, indeed, primarily a test of Newtonian beliefs about force and motion. The subsequent principal dimensions are considerably weaker, especially those in higher dimensions, and generally load on uninterpretably many distractors.

We were surprised by both the large fraction of shared sparse dimensions and their high degree of similarity (obvious in Figure 4) across Entropy, Geomin, and Bigeomin rotations. This is a testament to the invariance of the fundamental structure of the underlying data, since these rotations emphasize sparseness in substantially different ways. In cases where these best-matched dimensions do differ substantially, the alternate version is sometimes found in the local-optimum solutions of the other rotation methods; for example, D2 of geomin-0 does not match D2 of bigeominQ-0, but *does* closely match D2 of the second-best local optimum of the bigeomin criterion (bigeominQ-2; cf. Figure 5). Interestingly, *none* of the rotations find dimensions involving large clusters of distractors corresponding to Newton's three laws of motion, suggesting that such higher-level categorizations of mechanics are not monolithic in students' minds but are instead composed of several more atomic concepts with separately-measurable understanding.

 doi:10.20944/preprints202605.0952.v1

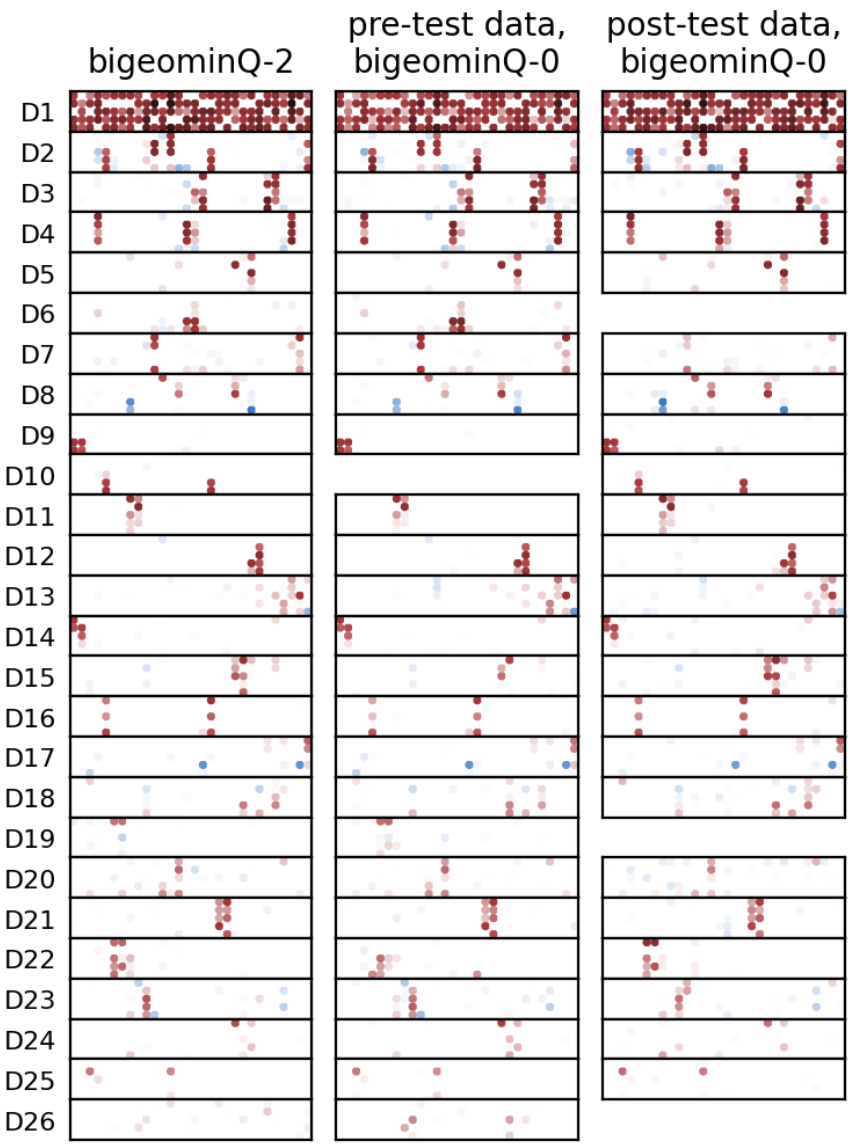

Figure 5. Sparse dimensions on all data, only pre-test data, and only post-test data. The dimensions are notably similar. Data subsets, by virtue of being smaller, produced fewer reliable dimensions, so gaps in the plot are due to no corresponding dimension being found when fitting to the subset.

From the various local solutions found by our different rotations, we have selected a subset of *Curated Sparse Dimensions*, emphasizing those that appeared consistently across different methods and initializations and/or whose set of prominent distractors seemed most likely to correspond to an interpretable psychological construct. We also leveraged information from a small bootstrap reproducibility study (described below) in making these choices, and favored exemplars from bigeominQ-0 when all other criteria did not strongly favor one specific solution.

The set of 29 curated dimensions along with associated statistics are tabulated in Table 2. The majority of these align closely with earlier results obtained from a smaller ∼34k-administration dataset ([17]); we name these using the same roman numerals as in that work for consistency, and also provide a description of what misconception they appear to encode based on our earlier work. For the two dimensions which lack an obvious "anti-Newtonian" sign convention (XI and XVIII), we additionally retain the same (arbitrary) sign convention as used in our previous work.

 doi:10.20944/preprints202605.0952.v1

4.2. Sparse Dimensions are Universal Patterns

We now demonstrate the universality of our results by showing their high degree of invariance across various subsets of the data, including:

- Random resamples of our complete dataset (via non-parametric bootstrapping)
- Administrations before and after instruction
- Students in our earlier 34k-administration dataset—over half of which was obtained from high-school students whose teachers had taken a two-week course in modeling instruction—and those in a completely non-overlapping subset of our current data representing mostly college students

4.2.1. Non-Parametric Bootstrap

As a first measure of overall reproducibility, we conducted a small bootstrap study by randomly generating 50 full-size bootstrap resamples of our 181k dataset, with sampling stratified to maintain a consistent ratio of pre- and post-test data. We ran our full hierarchical Bayes analysis on all such resamples as well as on the pre-test only and post-test only subsets of each, resulting in 150 sets of principal dimensions after inference. Each of our curated dimensions was then compared to the bootstrapped sets of principal dimensions using a form of Procrustes analysis: each principal solution was rotated obliquely to maximize the correlation coefficient with the curated dimension in question, and this "best aligned" correlation recorded. Recognizing that such a procedure would yield inflated correlations (as even a random matrix results will yield substantial correlations when rotated to maximize alignment in 26-dimensional space), we also computed random-chance baseline correlations using a permutation method, using these to linearly deflate each Procrustes-rotated bootstrap correlation ($r_{\text{adjusted}} = [r_{\text{original}} - r_{\text{chance}}] \,/\, [1 - r_{\text{chance}}]$). These results are reported in Table 2.

doi:10.20944/preprints202605.0952.v1

Table 2. Curated sparse dimensions. Note that names are not continuous. For each, we indicate the rotated solution set from which it was taken as well as the dimension number within that set (cf. Figure 4), the median bootstrap correlations in the complete dataset as well as in the pre- and post-test only subsets, and the sum of within-item variances of the dimension's $a^{(i,c)}$ parameters (which provide a rough overall measure of overall strength). For dimensions closely matching those found in our earlier work ([17]), we adopt the same roman-numeral designations and also note what physics misconception they appear to be encoding. We further illustrate the stability of these re-discovered results by providing uncentered correlation coefficients between the earlier dimensions (found from a much smaller dataset of ∼34k administrations) and those obtained from a completely independent (non-overlapping) subset of our current data.

		r-bootstrap					
Name	Source	all	pre	post	∑Var	Description	*r*-34k
I	bgQ-0 D1	1.00	0.99	0.98	23.8	Newtonian incorrect	0.99
II	bgQ-0 D4	1.00	0.98	0.98	2.6	Applied force exceeds resistance for constant speed	0.97
III	bgQ-0 D8	0.99	0.94	0.96	2.5	Impetus force, mostly on circular path	0.94
IV	bgQ-2 D8	0.99	0.91	0.94	1.1	Impetus plus centrifugal or centripetal force on circular path	0.93
V	bgQ-0 D3	1.00	0.98	0.98	2.4	Active dominates Massive dominates Passive	0.98
VI	bgQ-0 D2	0.99	0.94	0.94	2.1	Impetus force, mostly along linear path	0.93
VII	bgQ-2 D2	1.00	0.98	0.98	3.6	Impetus force on both linear and circular paths	0.98
VIII	bgQ-0 D6	0.99	0.97	0.70	1.2	Passive object pushed "because in the way"	0.92
IX	bgQ-0 D5	0.99	0.98	0.96	1.3	After rocket starts or stops, goes in direction of latest force or earlier motion	0.96
X	bgQ-0 D9	0.99	0.97	0.70	1.1	Omission of reaction force (normal or from passive object), usually with gravity	0.94
XI	bgQ-0 D12	0.99	0.94	0.94	1.1	Straight path preferred vs. curved paths	0.86
XIII	bgQ-0 D7	0.99	0.98	0.96	1.0	Decelerating after impulse from rocket firing	0.96
XIV	bgQ-0 D13	0.99	0.96	0.93	0.9	2M and M balls differ by factor of 2	0.88
XV	bgQ-0 D11	0.99	0.98	0.96	1.0	Latest force dominates after sudden force	0.93
XVI	bgQ-0 D21	0.98	0.86	0.90	0.7	Force change increases speed to a constant speed (not forever)	0.89
XVII	bgQ-1 D22	0.98	0.89	0.95	0.7	Continues curving inwards after inward force removed	0.92
XVIII	bgQ-0 D20	0.99	0.93	0.89	0.7	Air exerts significant drag and downward force	0.92
XIX	bgQ-0 D14	0.99	0.95	0.94	0.9	No acceleration due to rocket firing	0.87
XXI	bgQ-0 D10	0.99	0.97	0.95	1.1	2M ball falls significantly faster, but not by factor of 2	0.93
XXIII	bgQ-0 D19	0.98	0.93	0.91	0.9	Missing centripetal force	0.90
XXVI	bgQ-0 D16	0.99	0.97	0.91	0.6	Motion diagrams, v and a confused	0.94
XXVII	bgQ-0 D25	0.97	0.91	0.80	0.4	Gravity stronger closer to ground	0.81
XXVIII	bgQ-0 D15	0.98	0.91	0.92	0.6	–	–
XXIX	bgQ-0 D22	0.97	0.85	0.79	0.7	–	–
XXX	bgQ-1 D15	0.96	0.87	0.68	0.5	–	–
XXXI	bgQ-0 D24	0.97	0.90	0.81	0.5	–	–
XXXII	bgQ-0 D18	0.98	0.93	0.81	0.9	–	–
XXXIII	bgQ-0 D17	0.97	0.90	0.83	0.7	–	–
XXXIV	bgQ-0 D23	0.98	0.94	0.85	0.6	–	–

 doi:10.20944/preprints202605.0952.v1

Even after adjusting for chance, the median correlations for the full-sized bootstrap resamples are extremely high: the *minimum* adjusted value across all curated dimensions is 0.96, with most being 0.99 or greater. Those comparing the curated dimensions to pre- and post-only resample results are necessarily lower since cutting the size of the data in half leads to fewer dimensions being recovered during inference. Nevertheless, the majority remain high on at least one of the pre- or post-test results, with only two of our newly-identified dimensions (XXIX and XXX) falling below 90% on both. Overall, these results provide evidence that our sparse dimensions are quite invariant across student samples, with most also demonstrating invariance across pre- and post-testing.

4.2.2. Comparison of Pre- and Post-test Results

In addition to the adjusted bootstrap correlations above, we also explore the degree of similarity between sparse solutions obtained separately from the pre- and post-test data. These solutions were computed using the same methodology as our full-data results: application of the hierarchical Bayesian MNCM followed by rotation with several methods from 100 random initializations. We choose the best solutions obtained by the oblique bigeomin method (i.e., bigeominQ-0) in both cases, and present these results side-by-side in Figure 5. For comparison, we also include a bigeomin solution for the complete data; here we use one of our prominent local optima (bigeominQ-2) as this corresponds better to the solutions shown for the pre- and post-test solutions.

Clearly, an entire semester (or quarter) of Newtonian physics instruction does not significantly change the *nature* of most sparse dimensions. Of course, the *extent* to which each underlying misconception is held by a class of students will (hopefully) be reduced by instruction—this will be discussed in future work.

Despite the overall similarity, there are some notable differences between the three sets of results shown. First, the pre- and post-test analyses recover fewer overall dimensions, with D10 missing in the pre-test results and D6, D19, and D26 missing in the post-test results. This does not on its own indicate that these dimensions are *entirely* absent in the true generating model for these datasets—however, it does suggest that they are weak enough to be below the threshold of detectability given the smaller sample sizes and (likely) differing variance of the associated student traits across these datasets. Second, a small subset of dimensions are present but substantially altered in one of the solutions (e.g., D15 and D22 for the pre-test, D13 and D23 for the post-test). Further study is needed to determine whether these differences correspond to actual changes in students' underlying conceptual models or whether they are merely due to the sensitivity of the rotation minima to small perturbations of the loadings in such a high-dimensional space.

4.2.3. Comparison with Earlier Findings

As a final test, we compared sparse dimensions obtained from two non-overlapping subsets of our data with qualitatively different student demographics. The first of these subsets contains ∼34k administrations of the FCI, with more than half of these coming from high school students taught by teachers who had taken the Modeling Instruction workshop run by D. Hestenes' group at Arizona State University. Sparse dimensions found using these data were reported in previous work by our group (Segado, Adair, Stewart, and Pritchard in [17]), where they were found to correspond to identifiable misconceptions in physics known from PER literature and teacher experience.

The second subset contains only administrations that were *not* present in the original 34k dataset; it contains 148,473 administrations and consists primarily of the PhysPort and LASSO datasets in Table 1. Over 80% of these data were obtained from introductory college students who were taught with a variety of pedagogies, but rarely with modeling pedagogy.

The high degree of similarity between the "misconception dimensions" from our earlier study and the sparse dimensions obtained from non-overlapping data in the present work is immediately apparent in Figure 6. We quantify this similarity between the old dimensions and their closest analog in the new results using the uncentered Pearson correlation of the loadings, and include these values in the right-most column of Table 2 (labeled "r-34k"). All but four correlations are greater than 0.9, and

 doi:10.20944/preprints202605.0952.v1

none are below 0.8. The fact that so many of these sparse dimensions are so similar despite substantial differences in the demographics of the students from whom they are discovered is a very encouraging finding.

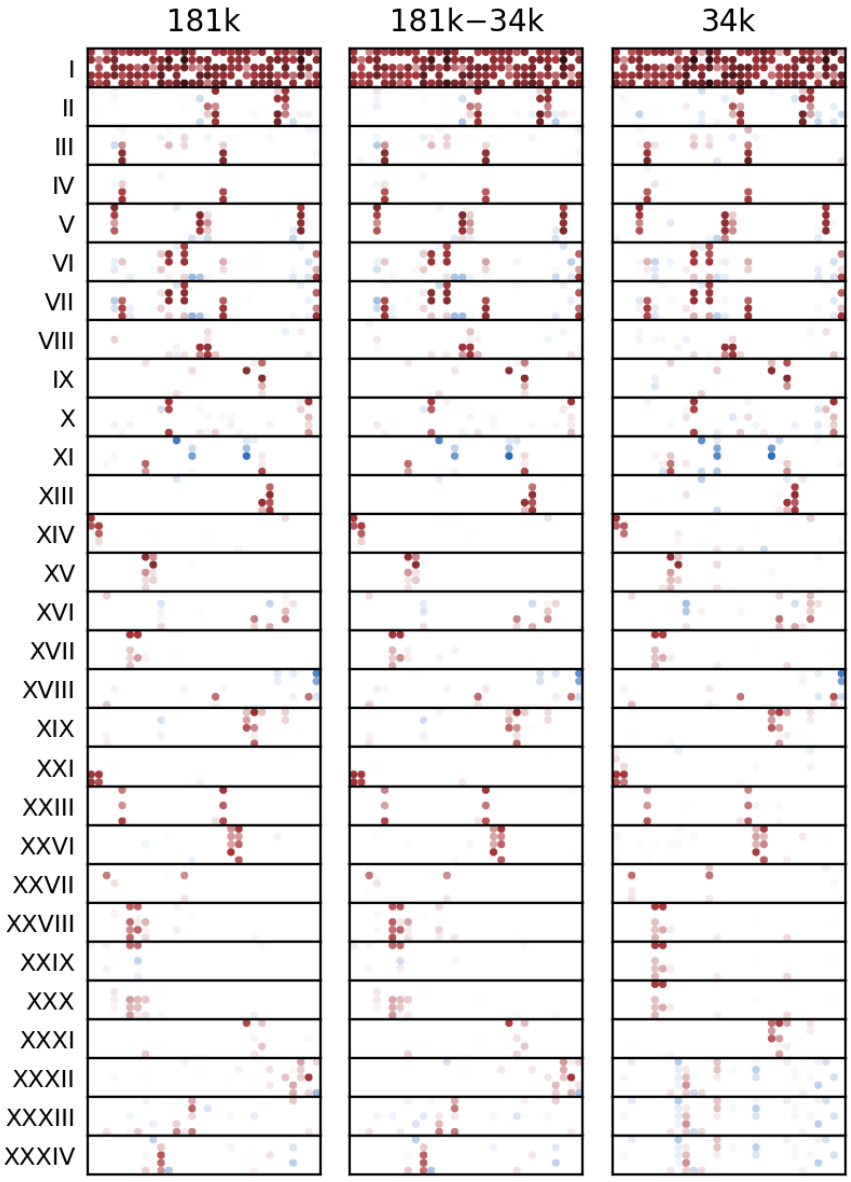

Figure 6. Curated sparse dimensions (left column) alongside similar dimensions obtained from two smaller datasets. The results labeled "34k" (right column; previously reported by Segado, 2025) were based on a much smaller dataset, over 50% of which consisted of high-school administrations of the FCI. Our complete 181k dataset includes these earlier data as well as a large number of new administrations from college students. The center column (labeled "181k−34k") illustrates results obtained from the relative complement of these two datasets—that is, from *only* the new administrations with none of the earlier 34k data included. Both the curated dimensions and the set obtained from new data alone bear strong similarity to dimensions in our earlier work, with more dimensions discoverable and higher dimensions clearer in the larger datasets.

Figure 6 also includes one 34k dimension which we had discarded in our earlier results as an uninterpretable residual (shown in the lower three panes of the right-most column). We observe that our larger dataset appears to resolve this into three distinct dimensions, all of which are simpler and—in our view—far more likely to admit meaningful interpretations.

 doi:10.20944/preprints202605.0952.v1

4.3. Most Curated Sparse Dimensions are Misconceptions.

Where possible, it is valuable to identify our sparse dimensions as specific misconceptions for three main reasons:

- It shows that these dimensions meet the definition of a stable mental misconception, thereby confirming the fundamental hypothesis of this paper
- It demonstrates the value of applying hierarchical Bayesian analysis methods in education
- It provides a list of validated "misconception dimensions" as a basis for future education-relevant work by us and others, and especially for those studying the FCI for whom replicating the full inference pipeline is not practical

As described above, most of our dimensions were already found to correspond to new or established misconceptions in earlier work, and a cursory examination reveals that these meanings are largely unchanged.

While we consider it important to interpret the remaining sparse dimensions in the hopes of classifying them as misconceptions (and have already formed several hypotheses), this is beyond the scope of this paper and will instead be addressed in a forthcoming one.

5. Summary and Future

We have demonstrated that application of hierarchical Bayesian methods leads to robust, reproducible solutions of the Multi-dimensional Nominal Categories IRT Model, even when the dimensionality of the recovered solutions is quite high. We also leverage our previous work in this area to conclude that most such solutions have meaningful psychological interpretation as misconceptions.

Along the way we have developed methods to recognize and eliminate multiple-choice response vectors from students who do not appear to respond conscientiously to the concept test we are analyzing, or who otherwise seem to base a substantial fraction of responses on guessing heuristics. This work can be viewed as the culmination of a series of studies on the FCI starting with dichotomous (right/wrong) factor analysis ([18,19]) on ∼21k-administration datasets, multi-dimensional models with ordered distractors ([20,21]), module analysis of response clusters ([22,23]), and finally the current Bayesian MNCM-based study that explores multidimensional information present in all distractors while leveraging a dataset that exceeded the size of previous ones by a factor of six.

Our table of curated sparse dimensions opens several important avenues for followup education-related work by our group. In addition to the obvious next step of providing interpretations for the new dimensions identified in the present work, we are also interested in seeing if dimensions encoding similar misconceptions will be found from another widely administered concept test in mechanics, the Force and Motion Concept Evaluation (FMCE). We have also begun exploring the score-dependent prevalence of each misconception for classes of students as well as the normalized gains of the same, which will enable us to compare their observed characteristics with previous research on misconceptions. This work will be published in more educationally- and cognitively-oriented journals.

Finally, the techniques described in this paper should be applicable to finding misconceptions from large-scale administrations of research-developed multiple choice instruments in other domains (provided their distractors reflect common student errors), or more broadly, to finding interpretable multidimensional representations of nominal data from nearly any source.

Author Contributions: Conceptualization: M.S., D.P., A.A; Data curation: A.D., M.S.; Formal analysis: M.S., A.A., Funding acquisition: D.P.; Methodology: M.S., D.P., A.A.; Software: M.S., A.D.; Supervision: D.P.; Writing: D.P., M.S., A.A., A.D., M.D. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Massachusetts Institute of Technology. This research received no external funding.

Institutional Review Board Statement: The MIT Committee on Use of Humans as Experimental Subjects has ruled that this work is exempt because it involves de-identified data only.

 doi:10.20944/preprints202605.0952.v1

Data Availability Statement: The datasets analyzed in this study are 3rd Party Data and were obtained from multiple independent sources and are not publicly archived due to privacy and institutional restrictions. Researchers seeking access to specific datasets may contact the respective data custodians directly: PhysPort data (smckagan@aapt.org); LASSO (Learning about STEM Student Outcomes, contact@LASSOeducation.org); Modeling data, (jane.jackson@asu.edu); Southwest and Central data (jcstewart1@mail.wvu.edu); Western Community College data (lhsu@santarosa.edu); Modeling data (American Modeling Teachers Association, engage@modelinginstruction.org); Eastern Private data (ms5629@drexel.edu); Misc. data (fuchsa@vaniercollege.qc.ca). For other queries regarding data used in this study, contact relatemit@gmail.com.

Acknowledgments: The 34k-dataset results presented in this work made use of computational resources provided by subMIT at MIT Physics. We thank John Stewart for additional help in the conceptualization of this work.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Baker, F.B. *The Basics of Item Response Theory*, 2nd ed.; ERIC Clearinghouse on Assessment and Evaluation: College Park, MD, 2001.
2. Segado, M.; Adair, A.; Stewart, J.; Ma, Y.; Drury, B.; Pritchard, D. A Multidimensional Bayesian IRT Method for Discovering Misconceptions from Concept Test Data. *Frontiers in Psychology* **2025**, *16*. https://doi.org/10.3389/fpsyg.2025.1506320.
3. Hestenes, D.; Wells, M.; Swackhamer, G. Force concept inventory. *The Physics Teacher* **1992**, *30*, 141–158. https://doi.org/10.1119/1.2343497.
4. Halloun, I.; Hake, R.; Mosca, E.; Hestenes, D. Force Concept Inventory, revised version (v95). Available at: https://www.physport.org/assessments/FCI, 1995.
5. Hake, R.R. Interactive-Engagement versus Traditional Methods: A Six-Thousand-Student Survey of Mechanics Test Data for Introductory Physics Courses. *American Journal of Physics* **1998**, *66*, 64–74. https://doi.org/10.1119/1.18809.
6. Coletta, V.P.; Phillips, J.A. Interpreting FCI scores: Normalized gain, preinstruction scores, and scientific reasoning ability. *American Journal of Physics* **2005**, *73*, 1172–1182, [https://pubs.aip.org/aapt/ajp/article-pdf/73/12/1172/13083783/1172_1_online.pdf]. https://doi.org/10.1119/1.2117109.
7. Savinainen, A.; Scott, P. Using the Force Concept Inventory to monitor student learning and to plan teaching. *Physics Education* **2002**, *37*, 53. https://doi.org/10.1088/0031-9120/37/1/307.
8. Takane, Y.; de Leeuw, J. On the Relationship between Item Response Theory and Factor Analysis of Discretized Variables. *Psychometrika* **1987**, *52*, 393–408. https://doi.org/10.1007/BF02294363.
9. Thissen, D.; Cai, L. Nominal Categories Models. In *Handbook of Item Response Theory*; Chapman and Hall/CRC, 2016.
10. Bock, D.R. Estimating Item parameters and Latent Ability when Responses Are Scored in Two or More Nominal Categories. *Psychometrika* **1972**, *37*, 29–51.
11. Falk, C.F.; Ju, U. Estimation of Response Styles Using the Multidimensional Nominal Response Model: A Tutorial and Comparison With Sum Scores. *Frontiers in Psychology* **2020**, *Volume 11 - 2020*. https://doi.org/10.3389/fpsyg.2020.00072.
12. Revuelta, J.; Franco-Martínez, A.; Ximénez, C. Nominal Factor Analysis of Situational Judgment Tests: Evaluation of Latent Dimensionality and Factorial Invariance. *Educational and Psychological Measurement* **2021**, *81*, 1054–1088, [https://doi.org/10.1177/0013164421994321]. PMID: 34565816, https://doi.org/10.1177/0013164421994321.
13. Revuelta, J. Multidimensional Item Response Model for Nominal Variables. *Applied Psychological Measurement* **2014**, *38*, 549–562, [https://doi.org/10.1177/0146621614536272]. https://doi.org/10.1177/0146621614536272.
14. Takane, Y. An Item Response Model for Multidimensional Analysis of Multiple-Choice Data. *Behaviormetrika* **1996**, *23*, 153–167. https://doi.org/10.2333/bhmk.23.153.
15. Natesan, P.; Nandakumar, R.; Minka, T.; Rubright, J.D. Bayesian Prior Choice in IRT Estimation Using MCMC and Variational Bayes. *Frontiers in Psychology* **2016**, *Volume 7 - 2016*. https://doi.org/10.3389/fpsyg.2016.01422.

 doi:10.20944/preprints202605.0952.v1

16. Zhang, H.; Chen, Y.; Li, X. A Note on Exploratory Item Factor Analysis by Singular Value Decomposition. *Psychometrika* **2020**, *85*, 358–372. https://doi.org/10.1007/s11336-020-09704-7.
17. Segado, M. Intuitive but Wrong: Uncovering Student Misconceptions About Force and Motion With Bayesian Item-Response Methods. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, United States, 2025.
18. Huffman, D.; Heller, P. What does the force concept inventory actually measure? *The Physics Teacher* **1995**, *33*, 138–143, [https://pubs.aip.org/aapt/pte/article-pdf/33/3/138/11752462/138_1_online.pdf]. https://doi.org/10.1119/1.2344171.
19. Eaton, P.; Willoughby, S.D. Confirmatory factor analysis applied to the Force Concept Inventory. *Phys. Rev. Phys. Educ. Res.* **2018**, *14*, 010124. https://doi.org/10.1103/PhysRevPhysEducRes.14.010124.
20. Stewart, J.; Zabriskie, C.; DeVore, S.; Stewart, G. Multidimensional item response theory and the Force Concept Inventory. *Phys. Rev. Phys. Educ. Res.* **2018**, *14*, 010137. https://doi.org/10.1103/PhysRevPhysEducRes.14.010137.
21. Stewart, J.; Drury, B.; Wells, J.; Adair, A.; Henderson, R.; Ma, Y.; Pérez-Lemonche, Á.; Pritchard, D. Examining the Relation of Correct Knowledge and Misconceptions Using the Nominal Response Model. *Physical Review Physics Education Research* **2021**, *17*, 010122. https://doi.org/10.1103/PhysRevPhysEducRes.17.010122.
22. Wells, J.; Henderson, R.; Stewart, J.; Stewart, G.; Yang, J.; Traxler, A. Exploring the structure of misconceptions in the Force Concept Inventory with modified module analysis. *Phys. Rev. Phys. Educ. Res.* **2019**, *15*, 020122. https://doi.org/10.1103/PhysRevPhysEducRes.15.020122.
23. Wells, J.; Henderson, R.; Traxler, A.; Miller, P.; Stewart, J. Exploring the structure of misconceptions in the Force and Motion Conceptual Evaluation with modified module analysis. *Phys. Rev. Phys. Educ. Res.* **2020**, *16*, 010121. https://doi.org/10.1103/PhysRevPhysEducRes.16.010121.