Variational Inference for the Partial Credit Model

Paolo Frazzetto[1,*], Nicolò Navarin[1] and Alessandro Sperduti[1,2]

[1]Department of Mathematics "Tullio Levi-Civita", University of Padova, 35121 Padua, Italy
[2]Augmented Intelligence Center, Bruno Kessler Foundation, 38123 Povo (TN), Italy

Abstract

Item Response Theory (IRT) models, particularly the Partial Credit Model (PCM), are indispensable in psychometrics for estimating unobserved latent abilities from questionnaires with ordered categorical responses. However, traditional estimation methods struggle with the scalability required for modern large-scale datasets. While Variational Autoencoders (VAEs) offer a promising path for scalable inference, their application to polytomous IRT models and the principled integration of respondent covariates remain underexplored. In this paper, we introduce a novel Variational Autoencoder framework for the Partial Credit Model (VA-PCM) that addresses these challenges. Our primary theoretical contribution is a psychometrically-informed generative model where the crucial ordering of item difficulty thresholds is guaranteed by construction. This is achieved by defining a Dirichlet prior over the proportions of the latent ability scale, which are then deterministically transformed into ordered thresholds via a stick-breaking process. Furthermore, we integrate respondent covariates as auxiliary variables whose distributions are explicitly conditioned on the latent abilities within a coherent probabilistic graphical model, allowing the model to leverage this information to enhance estimation. We present a complete proof-of-concept implementation using the Pyro probabilistic programming language, detailing the amortized inference architecture and deriving the Evidence Lower Bound (ELBO) for optimization. Preliminary experiments on synthetic data validate the framework's core mechanisms, demonstrating strong parameter recovery for the ordered item thresholds. The results also underscore the inherent difficulty of individual ability estimation from sparse response data, motivating clear directions for future work. By providing a complete theoretical and implementation blueprint, this work lays the groundwork for scalable, nuanced, and data-rich psychometric analysis of ordered response data.

Keywords

Psychometrics, Partial Credit Model, Variational Autoencoder, Ordinal Inference

1. Introduction

The quantitative measurement of unobserved human characteristics, such as cognitive abilities, personality traits, or attitudes, is a cornerstone of modern psychology, education, and the social sciences. Within this domain, Item Response Theory (IRT) has emerged as the dominant and most sophisticated paradigm for modeling the relationship between an individual's latent traits and their responses to a set of items on a questionnaire or test [1]. One of the most important models within the IRT family is the Partial Credit Model (PCM), which provides a framework for analyzing items with ordered, polytomous response categories, such as Likert scales or multi-step problems where partial credit is awarded [2, 3].

However, the advent of large-scale digital assessments, online learning platforms, and massive open online courses (MOOCs) has generated datasets of unprecedented size and complexity [4]. This explosion of data presents a significant computational challenge to traditional IRT estimation methods. Classical approaches like Marginal Maximum Likelihood (MML) estimation via the Expectation-Maximization (EM) algorithm become computationally intensive and may struggle with high-dimensional latent spaces, while Bayesian methods using Markov Chain Monte Carlo (MCMC) sampling, though robust, are often prohibitively slow for large datasets, hindering rapid model development and iteration [1, 5].

In response to these scalability challenges, recent research has turned to Variational Inference (VI), a technique from modern machine learning that reframes Bayesian inference as a fast and efficient optimization problem [6]. In particular, Variational Autoencoders (VAEs) have been successfully applied

ITADATA2025: The 4th Italian Conference on Big Data and Data Science, September 9–11, 2025, Turin, Italy

[*]Corresponding author.

paolo.frazzetto@unipd.it (P. Frazzetto); nicolo.navarin@unipd.it (N. Navarin); alessandro.sperduti@unipd.it (A. Sperduti)
0000-0002-3227-0019 (P. Frazzetto); 0000-0002-4108-1754 (N. Navarin); 0000-0002-8686-850X (A. Sperduti)

to dichotomous IRT models, demonstrating the ability to perform fast, scalable, and amortized inference, allowing for the instantaneous estimation of abilities for new individuals without retraining the model [1, 7, 4]. These works have established VAEs as a viable and powerful tool for large-scale psychometric analysis.

Despite these advances, the application of VAEs to the more complex Partial Credit Model has been less explored. Furthermore, most existing VAE-IRT frameworks do not provide a principled way to incorporate auxiliary respondent covariates (e.g., demographic or educational background), which can offer valuable information for improving the precision of ability estimates. This paper aims to fill these gaps by proposing a novel Variational Autoencoder framework specifically designed for the Partial Credit Model (VA-PCM). Our contributions are threefold: (1) we develop a generative model that faithfully represents the PCM for ordered categorical responses; (2) we integrate respondent covariates as an auxiliary source of information to enhance latent ability estimation within a coherent probabilistic graphical model; and (3) we formalize the corresponding amortized variational inference scheme and derive the Evidence Lower Bound (ELBO) for optimization.

This work lays the theoretical and methodological groundwork for applying deep generative modeling to polytomous response data at scale. Our preliminary experiments on synthetic data demonstrate the viability of the framework, showing its potential to recover item and person parameters, and setting the stage for future large-scale empirical validation. Lastly, the VA-PCM represents a step towards building more scalable, nuanced, and data-rich psychometric models for the modern era.

2. Background

2.1. Item Response Theory

Item Response Theory (IRT) encompasses a family of mathematical models that describe the probabilistic relationship between an individual's unobserved latent trait(s) and their observed responses to items on a test or questionnaire [1]. Unlike classical test theory, which focuses on aggregate test scores, IRT models the interaction at the item level, providing a more granular and theoretically grounded understanding of measurement.

The core idea of IRT is to characterize both respondents and items with a set of parameters on a common latent scale. A respondent is typically characterized by an ability parameter, denoted Z, which represents their level on the latent trait being measured. An item is characterized by one or more parameters, most commonly including a difficulty parameter (b), which indicates the ability level required for a 50% chance of a correct response, and a discrimination parameter (a), which reflects how well the item differentiates between individuals with different ability levels.

In its most common forms, such as the one-, two-, and three-parameter logistic models (1PL, 2PL, 3PL), IRT defines the probability of a correct response to a dichotomous (correct/incorrect) item using a logistic function. For example, the 2PL model is given by:

$$p(\text{correct}|Z, a, b) = \frac{1}{1 + e^{-a(Z-b)}}$$

This function produces a characteristic "S"-shaped curve where the probability of a correct response increases with the individual's ability Z. The primary task in IRT is *inference*: the estimation of these latent ability and item parameters from a matrix of observed responses.

While the foundational IRT models were developed for dichotomous data, many assessment contexts involve responses that are not simply right or wrong but reflect varying degrees of proficiency or endorsement. To handle such data, IRT was extended to polytomous models, which are designed for items with multiple, ordered response categories. It is within this class of models that the Partial Credit Model resides, providing a powerful tool for extracting nuanced information from ordered response data.

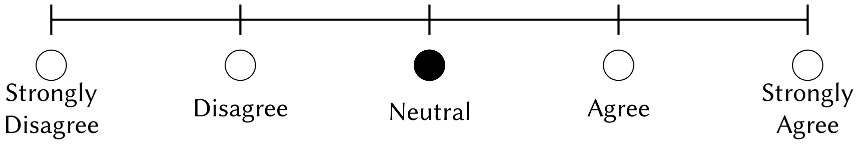

Figure 1: Example of a Likert scale with 5 values, which can be modeled with the PCM as one item of $m = 5$ categories. A respondent choose the "Neutral" answer, meaning $R_{n1} = 3$.

2.2. The Partial Credit Model

The Partial Credit Model (PCM), first introduced by [2] and formalized by [8], is a fundamental psychometric model within IRT designed for analyzing responses to items with ordered categorical scores. Unlike dichotomous models (e.g., the Rasch model) that classify responses as simply correct or incorrect, the PCM accounts for situations where responses can reflect varying degrees of correctness or proficiency, providing more nuanced insights into an individual's performance or attitude. This makes it particularly suitable for questionnaire items where responses are graded (e.g., essay scores, Likert scales, multi-step problem-solving tasks) or when multiple-choice items allow for partial credit (i.e., fractional scoring). See Figure 1 for an example of such scale.

The model is defined by a set of item parameters, which are the item's inherent characteristics. These characteristics are typically represented by a set of "step difficulty" or "threshold" parameters for each item. Generally speaking, the PCM is an ordered categorical model, where the logistic function of the difference between the respondent's latent ability and the item's difficulty gives the probability of a response.

Let N denote the number of respondents and I be the number of items in a questionnaire or test. For each respondent n and item i:

- Z_n represents the unobserved latent ability (trait) of respondent n. This ability is assumed to lie on a continuous scale.
- R_{ni} is the observed response r of respondent n to item i on—without loss of generality—a scale of integer values $\{0, \dots, m_i\} \ni r$ where m_i is the maximum possible score for item i. These categories are assumed to be ordered, reflecting increasing levels of proficiency. In the PCM, it is presumed that individuals with greater abilities tend to achieve higher scores on a given item. Nonetheless, the PCM does not make any assumption that there is an underlying sequential step process to achieve a score. Thus, it does not necessitate that a respondent must successfully complete all tasks associated with lower score categories to attain success in higher-scoring tasks [3].
- Ψ_i denotes the set of item parameters for item i. Specifically, Ψ_{ik} is the k-th step difficulty (or threshold) parameter for item i, associated with moving from score category $k-1$ to $k \le m_i$. Ideally, these $\Psi_{i,k}$ values should be monotonically increasing for a given item (i.e., $\Psi_{i1} < \Psi_{i2} < \dots < \Psi_{im_i}$), indicating that it becomes progressively more difficult to achieve higher scores. This assumption is only required for a psychometric interpretation, as unordered thresholds Ψ_{ik} do not per se violate its mathematical formulation [9]. For identifiability of a PCM instance, the propensity for the base category (score 0) is normalized to 1, which is equivalent to setting the first step difficulty $\Psi_{i0} = 0$ for all items. For simplicity, we assume that all items have the same amount of choices, so $m_i = m \; \forall \; i$.

The PCM is formally derived by applying the dichotomous Rasch model [10] to adjacent pairs of score categories for a polytomous item [2]. For an item with m categories, the probability of scoring k versus $k - 1$, given that the score is either $k - 1$ or k, is modeled as a dichotomous Rasch-like function:

$$p(R_{ni} = k | R_{ni} = k - 1 \text{ or } k, Z_n, \Psi_{ik}) = \frac{\exp(Z_n - \Psi_{ik})}{1 + \exp(Z_n - \Psi_{ik})}$$

This formulation highlights that each Ψ_{ik} can be interpreted as the ability level at which a respondent has a 50% chance of scoring k rather than $k-1$, when considering only these two adjacent categories. The

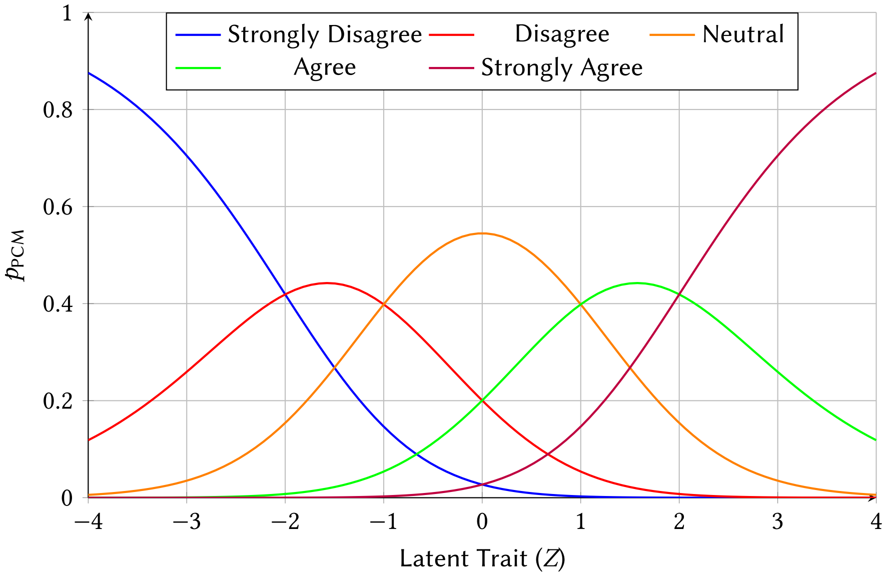

Figure 2: Category probability curves for a 5-point Likert item under the PCM. Each curve represents the probability of endorsing a specific response category as a function of the respondent's latent trait (Z). The PCM uses step parameters $\Psi = \{0, -2, -1, 1, 2\}$, where $\Psi_0 = 0$ by definition and the remaining values are increasing. As Z increases, the likelihood of selecting higher response categories increases consequently. Intersection points between adjacent curves indicate category thresholds, making this plot a useful tool in IRT for evaluating item functioning and category discrimination.

overall PCM probabilities for each score category are then derived from these conditional probabilities; hence, the probability of respondent n achieving a score r on item i according to the PCM is given by:

$$p_{\text{PCM}}(R_{ni} = r | Z_n, \Psi_i) = \frac{\exp(\sum_{k=0}^{r}(Z_n - \Psi_{ik}))}{\sum_{h=0}^{m} \exp(\sum_{k=0}^{h}(Z_n - \Psi_{ik}))} \tag{1}$$

where $r \in \{0, \dots, m\}$. The numerator represents the cumulative "propensity" for achieving score r or higher, while the denominator serves as a normalization constant, summing over all possible score categories for item i. Figure 2 depicts this function for a 5-point scale. From a computational perspective, this formulation essentially applies a softmax-like function to the cumulative log-odds of attaining each score category, with Ψ_{ik} serving as the ordered thresholds (or cutpoints) along the latent ability scale. This aligns with approaches used in probabilistic programming for ordinal regression, where observed categories are drawn from a categorical distribution whose logits are derived from these structured cumulative probabilities [11].

Boarder details about the PCM, its variations, interpretations, and parameters'meaning are outside the scope of this work and are deferred to the literature [3, 12, 9].

3. Variational Inference for the PCM

In real-world scenarios, given a $N \times I$ matrix of observed responses, we want to infer the ability of all N people and the characteristics of all I items. Next, we provide a brief overview of inference in IRT.

Traditional estimation methods for IRT models, such as Maximum Marginal Likelihood Estimation via the Expectation-Maximization (EM) algorithm [13] or Markov Chain Monte Carlo (MCMC) methods [14], have been foundational in psychometrics. However, these approaches face significant computational challenges when applied to large-scale datasets, complex model structures (e.g., high-dimensional latent spaces), or polytomous response formats like the PCM. The numerical integration required for marginalization in EM can become intractable, and MCMC, while providing robust posterior estimates, can be prohibitively slow, especially for estimating the large number of parameters involved in large-scale assessments [15].

Variational Inference (VI) has emerged from the machine learning and statistical communities as a powerful alternative for approximate Bayesian inference [6]. VI reframes inference as an optimization problem, seeking a tractable distribution that best approximates the true, often intractable, posterior distribution of latent variables. A key advantage of VI, particularly when coupled with amortized inference (as seen in VAEs [16]), is its remarkable speed and scalability to massive datasets. This efficiency is achieved by training neural networks to directly map observed data to the parameters of the approximate posterior, making subsequent inference queries highly efficient. While VAE-based approaches have shown promise for dichotomous IRT models [1], their extension to the PCM, which inherently models ordered categorical responses with multiple item thresholds, presents novel challenges and opportunities.

3.1. Integrating Respondent Covariates into the Generative Model

Our approach extends the PCM within a VAE framework to leverage auxiliary respondent information. In psychometric applications, additional data about respondents, such as age, education level, socio-economic status, or gender, are frequently available. These *respondent covariates* X can provide valuable insights into the latent abilities Z and enhance the precision of estimation. Rather than treating these covariates as mere descriptive statistics, our generative model integrates them directly, assuming that these *observed* features are, in part, determined by the underlying *latent* abilities. The idea is that these covariates are spurious associations that link one's features with their answers. Mathematically, the probability distribution of X is:

$$p(X, Z) = p(X|Z)p(Z)$$

We can then define the relationships among these variables by means of a probabilistic graphical model [17], as in Figure 3. This graphical structure implies a joint probability distribution over all observed and latent variables, which can be factorized according to the local Markov properties:

$$p(R, X, Z, \Psi) = p(R|Z, \Psi)p(X|Z)p(Z)p(\Psi) \tag{2}$$

In this factorization:

- $p(R|Z, \Psi)$ represents the PCM itself, describing the likelihood of observing a particular response R_{ni} given the respondent's latent ability Z_n and the item parameters Ψ_i, as previously defined in Eq. 1.
- $p(X|Z)$ models the relationship between the observed respondent features X_n and the latent abilities Z_n. This component allows the model to leverage rich covariate information to inform the estimation of Z_n. For instance, $p(X|Z)$ could be a simple linear model for continuous X or a logistic regression for categorical X.
- $p(Z)$ is the prior distribution over the latent abilities. A common choice is a standard Gaussian distribution, $\mathcal{N}(0, 1)$, reflecting an initial assumption of abilities centered around zero with unit variance.
- $p(\Psi)$ is the prior distribution over the item parameters $\Psi_i = (\Psi_{i0}, \ldots, \Psi_{i,m_i})$. Standard practice often uses independent Gaussian priors for each Ψ_{ik}. A key psychometric characteristic of the PCM is the expectation that these thresholds should be ordered. Standard independent Gaussian priors do not enforce this ordering, which can lead to ill-posed or uninterpretable thresholds during estimation. To address this, we employ a more principled prior based on a stick-breaking construction using the Dirichlet distribution [18]. Instead of defining a prior on the thresholds directly, we define a prior on the *proportions* $\Delta\Psi$ of the latent ability scale partitioned by the thresholds. A Dirichlet distribution, $\text{Dir}(\alpha)$, is a natural choice for this, as its samples are vectors of positive numbers that sum to one. We can then deterministically transform these proportions into a set of ordered thresholds Ψ_i on the logit scale, for instance by

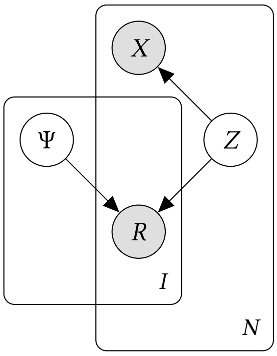

Figure 3: Probabilistic graphical model of a PCM with covariatex X, for I responses R given by N participants each with latent trait Z.

applying the inverse cumulative distribution function (CDF) of a standard normal distribution to the cumulative sums of the sampled proportions. This approach elegantly enforces the ordering constraint within the generative process itself, guiding the model towards more plausible and interpretable solutions.

The overall objective of the model is to learn parameters that maximize the likelihood of the observed data $\mathscr{L} = \log(p(R, X))$, which is obtained by marginalizing out the latent variables from the joint distribution:

$$p(R, X) = \int p(R, X, Z, \Psi) dZ d\Psi$$

which is intractable and needs to be approximated.

3.2. Variational Inference with Mean-Field Approximation

The Local Markov property states that each random variable in a random variable set $\mathscr{V}$ is conditionally independent of its non-descendants given its parent variables. We write:

$$X_v \perp\!\!\!\perp X_{\mathscr{V} \setminus \mathrm{de}(v)} \mid X_{\mathrm{pa}(v)} \quad \text{for all } v \in \mathscr{V}$$

where $\mathrm{de}(v)$ is the set of descendants, $\mathrm{pa}(v)$ is the set of parents, $\mathscr{V} \setminus \mathrm{de}(v)$ is the set of non-descendants of v. In our case, this can be expressed in the following relations:

$$Z \perp\!\!\!\perp \Psi;\ \Psi \perp\!\!\!\perp X, Z;\ X \perp\!\!\!\perp R, \Psi \mid Z;\ R \perp\!\!\!\perp X \mid \Psi, Z$$

In particular, from the latter two, it is clear that:

$$p(X|R, \Psi, Z) = p(X|Z) \tag{3}$$

$$p(R|X, \Psi, Z) = p(R|\Psi, Z) \tag{4}$$

In Variational Inference, instead of directly computing the intractable true posterior $p(Z, \Psi|R, X)$, we introduce a simpler, tractable variational distribution $q(Z, \Psi|R, X)$ that approximates it. The goal is to find the q distribution that is "closest" to the true posterior, typically measured by minimizing the Kullback-Leibler (KL) divergence $\mathrm{KL}(q(Z, \Psi|R, X) \| p(Z, \Psi|R, X))$. Minimizing the KL divergence is equivalent to maximizing the Evidence Lower Bound (ELBO) [6] which is a lower bound on the log marginal likelihood:

$$\mathscr{L}_{\mathrm{ELBO}}(q) = \log(p(R, X)) - \mathrm{KL}(q(Z, \Psi|R, X) \| p(Z, \Psi|R, X)). \tag{5}$$

To ensure tractability and enable efficient optimization, we typically employ a mean-field approximation for the variational distribution. This involves assuming that the approximate posterior factorizes over the latent variables in a so-called mean-field approximation, which for our model we define as:

$$q(Z, \Psi|R, X) = q_\Psi(\Psi|R) q_Z(Z|R, X), \tag{6}$$

where q_Ψ and q_Z are the approximate posteriors for the item parameters and latent abilities, respectively. The analogous form without covariates is $q(Z, \Psi|R) = q_\Psi(\Psi|R) q_Z(Z|R)$.

3.2.1. Amortized Inference and ELBO Optimization

For q_Z and q_Ψ, we choose flexible, tractable distributions whose parameters are determined by neural networks. This is known as *amortized inference* [19]. For $q_Z(Z|R, X)$, we can assume that the traits are normally distributed and then we can model it as a Gaussian distribution $\mathcal{N}(\mu_Z, \Sigma_Z)$, where the mean and covariance are outputs of a neural network with parameters θ_Z that takes R and X as inputs: $(\mu_Z, \Sigma_Z) = \mathrm{NN}_{\theta_Z}(R, X)$. For the item parameters $q_\Psi(\Psi|R)$, we mirror the structure of our prior. Instead of learning the parameters of a distribution over the thresholds directly, the inference network $\mathrm{NN}_{\theta_\Psi}$ learns to output the concentration parameters of a Dirichlet distribution for each item, $\alpha_i = \mathrm{NN}_{\theta_\Psi}(R_{\cdot i})$. The approximate posterior for the item-level proportions is thus $q(\Delta\Psi_i) = \mathrm{Dir}(\alpha_i)$. The final ordered thresholds Ψ_i are then obtained through the same deterministic transformation used in the generative model. This symmetric design ensures that our inference process respects the crucial ordering property of the PCM parameters, guiding the model towards psychometrically plausible solutions.

The parameters of these neural networks, θ_Z and θ_Ψ, are learned by maximizing the ELBO in Eq. 5 which becomes:

$$\begin{aligned}\mathscr{L}_{\mathrm{ELBO}}(q) &= \mathbb{E}_{q(Z,\Psi|R,X)}\left[\log p(R, X, Z, \Psi) - \log q(Z, \Psi|R, X)\right] \\ &= \mathbb{E}_{q_Z(Z|R,X)q_\Psi(\Psi|R,X)}\left[\log p(R|\Psi, Z) + \log p(X|Z) + \log p(Z) + \log p(\Psi)\right. \\ &\qquad\qquad\qquad\qquad \left. - \log q_Z(Z|R, X) - \log q_\Psi(\Psi|R, X)\right] \\ &= \mathbb{E}_{q_Z(Z|R,X)q_\Psi(\Psi|R,X)}\left[\log p(R|\Psi, Z)\right] + \mathbb{E}q_Z(Z|R, X)\left[\log p(X|Z)\right] \\ &\quad - \mathrm{KL}(q_Z(Z|R, X)\|p(Z)) - \mathrm{KL}(q_\Psi(\Psi|R, X)\|p(\Psi)) \end{aligned} \tag{7}$$

The ELBO consists of several terms with its own interpretation. The reconstruction term $\mathbb{E}_{q_Z(Z|R,X)q_\Psi(\Psi|R,X)}\left[\log p(R|\Psi, Z)\right]$ measures how well the latent variables and item parameters, sampled from their approximate posteriors, can reconstruct the observed responses and encourages accuracy in modeling the PCM. The auxiliary data likelihood term $\mathbb{E}q_Z(Z|R, X)\left[\log p(X|Z)\right]$ ensures that the latent abilities inferred from responses are consistent with the observed respondent features. This is how the auxiliary information X informs the latent ability estimates. Finally, the KL divergence terms $-\mathrm{KL}(q_Z(Z|R, X)\|p(Z)) - \mathrm{KL}(q_\Psi(\Psi|R, X)\|p(\Psi))$ act as regularization terms, encouraging the approximate posteriors to mimic their respective priors. Maximizing this ELBO with respect to the neural network parameters θ_Z and θ_Ψ is achieved through stochastic gradient descent.

4. Experiments and Implementation

To empirically validate our proposed Variational Autoencoder for the Partial Credit Model (VA-PCM), we conducted a series of experiments on synthetic data. The primary objectives of these experiments are to assess the model's ability to accurately recover the ground-truth parameters of the generative process and to detail the specific architectural and training methodologies required for a robust implementation. This section outlines the implementation details of our model, the hyperparameters used, and the design choices made to ensure stable and efficient training.

4.1. Implementation Details and Model Architecture

Our framework is implemented using PyTorch for neural network construction and gradient-based optimization, alongside Pyro [11], a deep probabilistic programming language, for defining the probabilistic model and performing variational inference.

4.1.1. Role of Respondent Covariates (X)

A key innovation of our framework is the principled integration of respondent covariates X. These covariates play a dual role in our methodology: first, as a fundamental component of the generative story that links latent abilities to observable characteristics, and second, as a critical source of information for the inference process.

In the Generative Process and Data Simulation Our generative model (Eq. 2) explicitly assumes that respondent covariates X are influenced by their latent abilities Z, captured by the conditional likelihood term $p(X|Z)$. To create synthetic data that faithfully adheres to this assumption, we generate covariates as a function of the ground-truth latent abilities. Specifically, we model this relationship as a linear transformation of the latent traits with additive Gaussian noise:

$$X_n = \beta Z_n + \epsilon_n, \quad \text{where } \epsilon_n \sim \mathcal{N}(0, \sigma_X^2)$$

Here, β is a weight matrix that defines the strength and nature of the relationship. This process ensures that the generated covariates X_n contain a quantifiable signal about the latent traits Z_n.

Within the VA-PCM's generative model, this relationship is parameterized by a dedicated decoder neural network NN_X. This network takes a sampled latent ability Z_n as input and outputs the parameters (mean and variance) of a Gaussian distribution for X_n. During training, the observed covariates are scored against this predicted distribution, contributing a likelihood term to the ELBO. This forces the model to learn a latent representation Z that is not only capable of explaining the observed responses R, but also the observed covariates X, thereby enforcing a powerful consistency constraint.

In the Inference Process The primary purpose of integrating covariates is to enhance the precision of latent ability estimation. This is achieved within the inference model, where the NN_Z network leverages the covariates as a direct input. The encoder's input for a given respondent n is a concatenation of their response vector R_n and their covariate vector X_n.

This design allows the inference network to fuse information from two distinct modalities: (1) behavioral response patterns captured in R, and (2) contextual or demographic features contained in X. By having access to both sources of evidence, the encoder can produce a more robust and precise estimate of the latent ability posterior, $q_Z(Z|R, X)$. For instance, if a respondent's answers are ambiguous, their covariates can provide an additional signal that helps to resolve the uncertainty in their estimated ability, leading to more accurate inference.

4.1.2. The Generative Model (Decoder)

The generative model, or *decoder*, programmatically defines the joint distribution outlined in Eq. 2. The priors for the latent variables are specified first. The latent abilities Z for each respondent are drawn from a standard normal prior, $p(Z) = \mathcal{N}(0, 1)$. For the item thresholds Ψ, we adopt the method described above to enforce the ordering constraint $\Psi_{i1} < \Psi_{i2} < \cdots < \Psi_{im_i}$, which is crucial for psychometric interpretability. Specifically, we sample a vector of proportions for each item from a Dirichlet prior, $p(\Delta\Psi_i) = \text{Dir}(\alpha)$, and then deterministically transform these proportions into a set of ordered thresholds on the logit scale using the inverse CDF of a standard normal distribution. This generative process guarantees that the sampled thresholds are correctly ordered by construction.

The likelihood functions for the observed data are conditioned on these latent variables. The relationship between covariates and abilities, $p(X|Z)$, is modeled by a dedicated decoder neural network NN_X, which takes a sampled latent ability Z_n as input and outputs the parameters (mean and variance) of a Gaussian distribution from which the observed covariates X_n are assumed to be drawn. The response likelihood, $p(R|Z, \Psi)$, is implemented according to the PCM formula (Eq. 1), where the logits for the categorical distribution of responses are computed from the sampled abilities Z and item thresholds Ψ.

4.1.3. The Inference Model (Encoder)

The inference model, or *encoder*, specifies the variational distribution $q(Z, \Psi|R, X)$ that approximates the true posterior. This is achieved through amortized inference, where neural networks learn to map observed data directly to the parameters of the approximate posterior distributions.

- **Ability Inference (q_Z):** The NN_Z network infers each respondent's latent ability. Its input is the concatenation of the respondent's covariate vector X_n and an embedding of their full response vector R_n. To handle the categorical nature of responses, each item's response is passed through a dedicated embedding layer before concatenation. The network then outputs the mean and log-variance of the Gaussian approximate posterior for Z_n.

- **Item Parameter Inference (q_Ψ):** The NN_Ψ network infers the parameters for each item. This network exemplifies one of the key complexities in modeling item-level parameters in an amortized fashion. Since item parameters are global, the network must summarize the information from all responses given to a specific item. For each item i, we compute a feature vector consisting of summary statistics (mean and standard deviation of responses) and the empirical distribution of response categories. This feature vector is then fed into the network to produce the concentration parameters of a Dirichlet distribution, which serves as the approximate posterior for the item's threshold proportions, $q(\Delta\Psi_i|R) = \text{Dir}(\alpha_i)$. This symmetric design, where the variational posterior for the proportions mirrors the Dirichlet prior, ensures that the inferred thresholds also adhere to the ordering constraint.

The training process involves optimizing the ELBO via Pyro's Stochastic Variational Inference (SVI) engine, as detailed in Algorithm 1. This framework handles the complexities of the model, such as computing gradients through stochastic latent variables (via the reparameterization trick) and managing the interplay between the probabilistic model and the deep neural networks.

4.2. Experimental Setup and Hyperparameters

Our experiments are conducted on synthetic data generated according to the process described in the previous section. This allows us to control the ground truth and systematically evaluate our model's performance.

4.2.1. Training Procedure

The model is trained end-to-end by maximizing the ELBO using mini-batch stochastic gradient descent. We employ the *ClippedAdam* optimizer [20], a variant of Adam with gradient clipping, which we found to provide additional stability during training. To enhance convergence and prevent common failure modes in VAEs, we incorporate several more best-practice techniques:

- **KL Annealing:** In the initial phases of training, the KL divergence term in the ELBO can overwhelm the reconstruction loss, causing the approximate posterior to collapse onto the prior (i.e., the "posterior collapse" problem). To mitigate this, we use KL annealing, where the KL terms are multiplied by a coefficient β that is gradually increased from a small initial value (e.g., $\beta = 0.01$) to its final value of $\beta = 1.0$ over a set number of training epochs. This allows the model to first focus on learning to reconstruct the data before being strongly regularized by the prior.
- **Learning Rate Scheduling:** We use a learning rate scheduler that gradually decays the learning rate after each epoch. This helps the optimizer to take smaller steps as it approaches a minimum, leading to finer convergence.
- **Early Stopping and Dropout:** To prevent overfitting and reduce unnecessary computation, we employ an early stopping (stopping training after 50 epochs without improvements) and dropout mechanism ($p_{dropout} = 0.1$).

4.2.2. Hyperparameter Configuration

The specific hyperparameters for our neural network architectures and training procedure were selected based on common practices for VAEs and preliminary experimentation. The key settings used in our experiments are summarized in Table 1.

Algorithm 1 VI-PCM Forward Pass

```
Require: Observed responses $R \in \mathbb{R}^{N \times I}$, Observed respondent features $X \in \mathbb{R}^{N \times \dim(X)}$
  Initialize neural networks: $\mathrm{NN}_Z$, $\mathrm{NN}_\Psi$, and the covariate decoder $\mathrm{NN}_X$.
  Initialize overall ELBO: $\mathscr{L}_{\mathrm{ELBO}} \leftarrow 0$
      {— Part 1: Infer Item Parameters from Responses —}
  Initialize item parameter KL divergence: $\mathrm{KL}_\Psi \leftarrow 0$
  for $i = 1 \dots I$ do
    Extract all responses for item $i$: $R_{\cdot i} = (R_{1i}, \dots, R_{Ni})$
    Compute variational parameters for item $i$'s threshold proportions: $\alpha_i = \mathrm{NN}_\Psi(R_{\cdot i})$
    Sample item proportions $\Delta\Psi_i \sim \mathrm{Dirichlet}(\alpha_i)$ {Sample from $q_\Psi$}
    Deterministically transform proportions $\Delta\Psi_i$ to ordered thresholds $\Psi_i$
    Compute KL term for item $i$: $\mathrm{KL}_{\Psi_i} = \mathrm{KL}(\mathrm{Dirichlet}(\alpha_i) || p(\Delta\Psi_i))$
    $\mathrm{KL}_\Psi \leftarrow \mathrm{KL}_\Psi + \mathrm{KL}_{\Psi_i}$
  end for
  $\mathscr{L}_{\mathrm{ELBO}} \leftarrow \mathscr{L}_{\mathrm{ELBO}} - \mathrm{KL}_\Psi$
      {— Part 2: Infer Latent Abilities and Compute Likelihoods for each Respondent —}
  for $n = 1 \dots N$ do
    Extract data for respondent $n$: Response vector $R_n$ and covariate vector $X_n$
      {Use inference network $\mathrm{NN}_Z$ to get parameters for $q_Z(Z_n | R_n, X_n)$}
    Compute variational parameters for latent ability $Z_n$: $(\mu_{Z_n}, \Sigma_{Z_n}) = \mathrm{NN}_Z(R_n, X_n)$
    Sample latent ability $Z_n \sim \mathcal{N}(\mu_{Z_n}, \Sigma_{Z_n})$ {Sample from $q_Z$}
      {Compute the three components of the ELBO for respondent $n$}
    Response Reconstruction Term:
    $\mathscr{L}_{\mathrm{Recon},n} = \sum_{i=1}^{I} \log p_{\mathrm{PCM}}(R_{ni} | Z_n, \Psi_i)$ {Using the sampled $Z_n$ and $\Psi_i$}
    Auxiliary Data Likelihood Term:
    Use the covariate decoder $\mathrm{NN}_X$ to parameterize the likelihood $p(X_n | Z_n)$
    $(\mu_{X_n}, \Sigma_{X_n}) = \mathrm{NN}_X(Z_n)$
    $\mathscr{L}_{\mathrm{Aux},n} = \log \mathcal{N}(X_n | \mu_{X_n}, \Sigma_{X_n})$ {Score observed $X_n$ under generated distribution}
    Latent Ability KL Divergence:
    $\mathrm{KL}_{Z,n} = \mathrm{KL}(\mathcal{N}(\mu_{Z_n}, \Sigma_{Z_n}) || p(Z_n))$
      {Update total ELBO with respondent's contribution}
    $\mathscr{L}_{\mathrm{ELBO}} \leftarrow \mathscr{L}_{\mathrm{ELBO}} + \mathscr{L}_{\mathrm{Recon},n} + \mathscr{L}_{\mathrm{Aux},n} - \mathrm{KL}_{Z,n}$
  end for
  return Return and Backpropagate on $\mathscr{L}_{\mathrm{ELBO}}$
```

4.3. Preliminary Results on Synthetic Data

To provide an initial proof-of-concept and assess the fundamental viability of our proposed VA-PCM framework, we conducted a preliminary experiment using synthetic data. The data were generated according to the process detailed in Section 4, with a configuration designed to represent a common, moderately-sized scenario: $N = 1000$ respondents, $I = 10$ items, a single latent dimension, a single covariate dimension, and $m = 5$ response categories per item.

After training the VA-PCM model on this dataset, we evaluated its ability to recover the ground-truth latent abilities Z and item threshold parameters Ψ. The results of this parameter recovery analysis are summarized in Table 2.

The analysis of the item parameter recovery is highly encouraging. We observe a strong positive correlation of 0.750 between the estimated and true item thresholds. This indicates that the model is successfully capturing the relative ordering and difficulty of the item steps. Furthermore, the R^2 score of 0.498 suggests that our model can account for approximately 50% of the variance in the true item parameters, a respectable result given the complexity of the model and the size of the dataset.

In contrast, the recovery of the respondent latent abilities presents a notable challenge in this initial

Table 1
Hyperparameter settings for the VA-PCM experiments.

Parameter	Value
Architecture	
Ability Encoder Hidden Dims	[64, 32]
Ability Encoder Embedding Dim	16
Threshold Encoder Hidden Dims	[32, 16]
Covariate Decoder Hidden Dims	16
Training Procedure	
Epochs	300
Batch Size	1000 (Full Dataset)
Learning Rate	0.0001
Learning Rate Decay	0.99
Weight Decay	1×10^{-4}
KL Annealing	$\beta = 0.05$

Table 2
Parameter recovery metrics for the VA-PCM on a preliminary synthetic dataset.

Parameter	Correlation	RMSE	R^2 Score
Ability (Z)	0.344	0.884	-0.007
Thresholds (Ψ)	0.750	0.742	0.498

experiment. The correlation between the estimated and true abilities is modest at 0.344, indicating a positive but weaker association. More significantly, the negative R^2 score of -0.007 reveals that the model's predictions for latent ability are, on average, less accurate than simply using the mean of the true abilities as a prediction. This highlights the inherent difficulty of person parameter estimation, particularly with shorter test lengths (only 10 items), where the amount of data available for any single respondent is limited. Nonetheless, these preliminary findings serve as a crucial benchmark and motivate more extensive experimental plans.

5. Conclusion and Future Work

In this paper, we have introduced a novel variational inference framework for the Partial Credit Model (VA-PCM), designed to address the scalability limitations of traditional psychometric estimation methods while providing a principled mechanism for integrating respondent covariates. We established a solid probabilistic framework, formalizing the relationships between responses, covariates, and latent variables within a generative graphical model. By leveraging amortized inference via neural networks and employing psychometrically-informed priors and variational families that respect the ordered nature of PCM parameters, we have outlined a complete and feasible methodology for applying modern deep generative modeling to polytomous response data.

Our preliminary experiments on synthetic data offer a promising, albeit mixed, proof-of-concept. The model demonstrates a strong capacity to recover the underlying structure of item parameters, which is a critical requirement for any psychometric model. However, the results also underscore the well-known challenges of accurately estimating individual-level abilities from limited data, a task that remains difficult even with the inclusion of covariates.

This work is primarily a theoretical and methodological contribution, intended to lay the groundwork for future research. We acknowledge that our experimental evaluation is preliminary. A comprehensive validation of the VA-PCM will require a more extensive set of experiments, which, while essential, are both computationally and temporally demanding. Key directions for future research include carrying out more experiments in different settings, with a particular focus on testing the model on larger datasets and under varying conditions of missing data. It is also essential to conduct a rigorous comparison of the VA-PCM against established psychometric baselines, such as MML-EM and MCMC-based methods, to benchmark its performance in terms of both accuracy and computational efficiency. Another important direction is the application of the framework to a large-scale, real-world dataset, such as PISA or TIMSS, to demonstrate its practical utility, scalability, and the interpretability of its findings in authentic educational contexts. Finally, given the flexibility of the proposed framework, future work could explore model extensions, such as incorporating multi-dimensional latent traits to capture more complex cognitive structures, or experimenting with advanced neural architectures and prior distributions to enhance estimation accuracy, such as normalizing flows.

Ultimately, this work aims to bridge the gap between traditional psychometrics and modern deep generative modeling. By demonstrating how the Partial Credit Model can be robustly integrated into a VAE framework, complete with covariates and principled handling of ordered data, we hope to foster further interdisciplinary research and pave the way for more scalable, nuanced, and data-rich psychometric analyses.

Acknowledgments

We would like to thank Prof. Barbara Hammer, Prof. Benjamin Paassen, and Amajor S.p.A. for the valuable support and inspiration to conduct this research. We acknowledge the support of the PNRR project FAIR - Future AI Research (PE00000013), Concession Decree No. 1555 of October 11, 2022, CUP C63C22000770006 and of the project "Future AI Research (FAIR) - Spoke 2 Integrative AI - Symbolic conditioning of Graph Generative Models (SymboliG)" funded by the European Union under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.3 - Call for tender No. 341 of March 15, 2022 of the Italian Ministry of University and Research – NextGenerationEU, Code PE0000013, Concession Decree No. 1555 of October 11, 2022, CUP C63C22000770006.

Declaration on Generative AI

The author(s) have employed Generative AI tools for proofreading and improving the readability of figures and tables.

References

[1] M. Wu, R. L. Davis, B. W. Domingue, C. Piech, N. Goodman, Variational item response theory: Fast, accurate, and expressive, arXiv preprint arXiv:2002.00276 (2020).

[2] G. N. Masters, A rasch model for partial credit scoring, Psychometrika 47 (1982) 149–174.

[3] M. Wu, H. P. Tam, T.-H. Jen, Partial Credit Model, in: M. Wu, H. P. Tam, T.-H. Jen (Eds.), Educational Measurement for Applied Researchers: Theory into Practice, Springer, Singapore, 2016, pp. 159–185. doi:10.1007/978-981-10-3302-5_9.

[4] M. Curi, G. A. Converse, J. Hajewski, S. Oliveira, Interpretable variational autoencoders for cognitive models, in: 2019 international joint conference on neural networks (ijcnn), IEEE, 2019, pp. 1–8.

[5] C. Cui, C. Wang, G. Xu, Variational estimation for multidimensional generalized partial credit model, psychometrika 89 (2024) 929–957.

[6] D. M. Blei, A. Kucukelbir, J. D. McAuliffe, Variational inference: A review for statisticians, Journal of the American statistical Association 112 (2017) 859–877.

[7] B. PaaBen, M. Dywel, M. Fleckenstein, N. Pinkwart, Sparse factor autoencoders for item response theory., International Educational Data Mining Society (2022).

[8] E. Muraki, A Generalized Partial Credit Model: Application of an EM Algorithm, Applied Psychological Measurement 16 (1992) 159–176. doi:10.1177/014662169201600206.

[9] R. J. Adams, M. L. Wu, M. Wilson, The rasch rating model and the disordered threshold controversy, Educational and Psychological Measurement 72 (2012) 547–573.

[10] G. Rasch, Probabilistic models for some intelligence and attainment tests., ERIC, 1993.

[11] E. Bingham, J. P. Chen, M. Jankowiak, F. Obermeyer, N. Pradhan, T. Karaletsos, R. Singh, P. Szerlip, P. Horsfall, N. D. Goodman, Pyro: Deep universal probabilistic programming, Journal of Machine Learning Research (2018). See also the ordinal regression tutorial: https://num.pyro.ai/en/stable/tutorials/ordinal_regression.html.

[12] R. J. Adams, M. L. Wu, The Mixed-Coefficients Multinomial Logit Model: A Generalized Form of the Rasch Model, in: Multivariate and Mixture Distribution Rasch Models: Extensions and Applications, Statistics for Social and Behavioral Sciences, Springer, 2007, pp. 57–75. doi:10.1007/978-0-387-49839-3_4.

[13] R. D. Bock, M. Aitkin, Marginal maximum likelihood estimation of item parameters: Application of an em algorithm, Psychometrika 46 (1981) 443–459.

[14] J.-S. Kim, D. M. Bolt, Estimating item response theory models using markov chain monte carlo methods, Educational Measurement: Issues and Practice 26 (2007) 38–51.

[15] L. Cai, A two-tier full-information item factor analysis model with applications, Psychometrika 75 (2010) 581–612.

[16] D. P. Kingma, S. Mohamed, D. Jimenez Rezende, M. Welling, Semi-supervised learning with deep generative models, Advances in neural information processing systems 27 (2014).

[17] D. Koller, N. Friedman, Probabilistic graphical models: principles and techniques, MIT press, 2009.

[18] M. Betancourt, Ordinal regression case study, 2019. URL: https://betanalpha.github.io/assets/case_studies/ordinal_regression.html, section 2.2.

[19] S. Gershman, N. Goodman, Amortized inference in probabilistic reasoning, in: Proceedings of the annual meeting of the cognitive science society, volume 36, 2014.

[20] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).