SPRITE: A Data-Driven Response Model For Multiple Choice Questions

Andrew E. Waters, Andrew S. Lan, Ryan Ning, Christoph Studer, and Richard G. Baraniuk

Abstract—Item response theory (IRT) models, in their most basic, dichotomous form, model a set of binary-valued (correct/incorrect) responses from individuals to items/questions. These models are ubiquitous in computer-based learning analytics and assessment applications, because they enable the inference of latent student abilities/respondent traits. Since the option a student selects on a multiple-choice question (either the correct response or one of the incorrect, distractor responses) contains more information regarding the student's ability than a simple binary-valued grade, polytomous IRT models have been developed to cover the cases of unordered (i.e., categorical) options and strictly ordered (i.e., ordinal) options. However, in many real-world educational scenarios, the various distractor options in a multiple-choice question are neither categorical, since they are incorrect to varying degrees, nor ordinal, since they are not strictly ordered. Moreover, this (partial) ordering information might not be known a priori, inhibiting the application of existing polytomous IRT models to practical scenarios. In this work, we propose the SPRITE (short for stochastic polytomous response item) model, a novel IRT extension for multiple-choice questions with unknown, partially ordered options. SPRITE improves substantially over existing IRT models in that it (i) learns the (partial) ordering of the options directly from student response data, (ii) produces interpretable model parameters, and (iii) outperforms existing approaches on predicting unobserved student responses on multiple real-world educational datasets.

Index Terms—Item response theory, learning analytics, categorical data, ordinal data

1 Introduction

Recent advances in machine learning and big data enable the delivery of personalized learning experiences tailored to each individual student at a massive scale by analyzing the rich data generated by students and using the results to close the learning feedback loop [7], [18], [34].

1.1 Item response theory

The first step to building a personalized learning system (PLS) is to create a statistical model that enables the analysis of student responses to questions—a ubiquitous source of data in any learning environment. The purpose of such a model is to evaluate how well the students have mastered some set of concepts/skills/knowledge components [2], [10], [18]. The simplest approach to evaluate the students' performance is to simply count the number of questions they answer correctly, referred to as the classical test theory (CTT) [3]. However, CTT ignores the detailed response patterns of each student, which can provide additional insights on the students' knowledge states. For example, two students can have the same number of correct responses yet have completely different areas of mastery. Such information that cannot be modeled by a simple aggregate score.

Item response theory (IRT) addresses this issue using a probabilistic model to characterize the response from each student to each question[1] in order to extract the ability of each student and the difficulty of each question [22]. The most basic form of IRT explicitly models the binary-valued (correct/incorrect) student responses to questions by characterizing the probability that a student will answer a question correctly as a function of the student's ability and the question's difficulty. IRT is widely considered to be superior than CTT, since it achieves higher precision in estimating a student's ability with a smaller number of questions compared to CTT [22]. Since its creation, IRT has seen wide-spread adoption in analyzing surveys, questionnaires, and standardized tests, including the Graduate Record Examination (GRE) and the Graduate Management Admission Test (GMAT) [36].

A student's response to a multiple-choice question, the most common type of question in educational assessments today, carries potentially more information into their ability than a binary-valued grade [31]. The reason is that as the specific mulitple-choice options (especially the incorrect options, which are often known as *distractors*) can be designed to not only indicate insufficient knowledge but also to reveal specific areas of deficiency and specific types of misconception [29]. As a result, there have been numerous extensions to the basic IRT model that model the probability of a student selecting each particular option in a multiple-choice question. These extensions can be classified into the following two types.

The first type of model treats the options as completely unordered, often referred to as *categorical* options. An example of a question with categorical options is the first question shown in Table 1. The use of such categorical models, notably the nominal

- *A. S. Lan, R. Ning, and R. G. Baraniuk are with the Department of Electrical and Computer Engineering, Rice University.*
- *A. E. Waters is with OpenStax.*
- *C. Studer is with the Department of Electrical and Computer Engineering, Cornell University.*

1. The term "item" in IRT can correspond to any item that requires respondents to respond to. In the education and testing domains, the term "item" corresponds to assessment questions in quizzes, homework assignments, or exams.

TABLE 1
Three examples of multiple-choice questions with categorical (unordered) options, ordinal (strictly ordered) options, and partially ordered options.

	Categorical	**Ordinal**	**Partially ordered**
Option	What is your favorite animal?	How many hours do you sleep per day?	What is the capital of Brazil?
A	Elephant	< 6	São Paulo
B	Eagle	$6 - 7$	Rio de Janeiro
C	Dolphin	$8 - 9$	Beijing
D	Platypus	> 9	Brasilia

response model (NRM) [6], produce student parameters that are not necessarily correlated with their abilities [30], which is critical to many applications where student ability estimates are needed to provide feedback on their strengths and weaknesses.

The second type of model treats the options as strictly ordered, often referred to as *ordinal* options, and assume that the ordering is known a priori. An example question with ordinal options is the second question in Table 1. Examples of these models include the generalized partial credit model (GPCM) [21], and the ordinal (ORD) model which relies on the assumption of a discrete set of ordered bins which, when combined with a latent ability variable, induce a probability distribution over the set of ordered options.

The strict and known ordering assumptions in both of the above ordinal models are often too restrictive in practice, since *many questions do not have strictly ordered options.* The third question in Table 1 provides an example of a question with partially ordered options. For this question, Option D is the correct answer; Options A and B are not correct but, since both are large cities in Brazil, can be considered to be equally incorrect; Option C would be considered the most incorrect option, since Beijing is not even in South America. Thus, for this question, a strict ordering of the options does not exist. We discuss the details of the NRM, GPCM, and ORD models in Section 2.

1.2 Limitations of existing models

As discussed above, most IRT models assume that the multiple-choice options are either categorical or ordinal with the ordering assumed to be known a priori. For most real-world educational scenarios, these assumptions are unrealistic. As we showed in the above example (the third question in Table 1) there might not be a strict ordering among the options, since there might be a correct option and multiple distractor options that are equally incorrect. Furthermore, the (partial) ordering of the options is often not known a priori unless a domain expert provides this information. This manual-labeling approach is labor-intensive and does not scale to large-scale educational scenarios with millions of students and thousands of questions.

In summary, there is a need for a new polytomous IRT model for multiple-choice questions that can deal with options that are partially ordered or have unknown ordering.

1.3 Contributions

We propose a novel IRT model for multiple-choice questions with unknown, and potentially partially ordered, options that we dub SPRITE (short for stochastic polytomous response item) model. Under these modeling assumptions, we provide an algorithm for SPRITE that automatically learns a (possibly partial) ordering among the options of each question solely from data. In addition, SPRITE provides a high level of interpretability and is able to provide statistics on the informativeness of questions and options.

Furthermore, SPRITE provides (often significantly) better prediction performance on unobserved student responses than existing IRT models. Table 2 demonstrates the superiority of SPRITE for a collection of real-world datasets in terms of the performance on predicting unobserved student responses (i.e., the options each student selects on each question in a held-out dataset). The details about the individual datasets are summarized in Table 3. For this experiment we train each model on 80% of the student responses in each dataset and use the estimated modeled parameters to predict the 20% holdout responses. We compute the prediction error rate as the ratio of the number of incorrect predictions over the total number of predictions and see that SPRITE outperforms the competing models on every dataset.

1.4 Paper outline

This paper is organized as follows. In Section 2, we review existing IRT models. In Section 3, we introduce the SPRITE model. In Section 4, we develop a Markov Chain Monte-Carlo (MCMC) sampling method for parameter inference under the SPRITE model. In Section 5, we present experimental results on both synthetic and real-world datasets. We conclude in Section 6.

2 Existing Statistical Models For IRT

We start by describing our notation and basic IRT models. We then discuss existing categorical and ordinal polytomous IRT models.

2.1 IRT notation and modeling assumptions

Assume that we have a dataset consisting of the responses from N students (or respondents, test takers) to Q questions (e.g., multiple-choice questions in an educational scenario). The observed data matrix $\mathbf{Y}$ consists of all the options each student selects on each question, with Y_{ij} denoting the option the i^{th} student selects on the j^{th} question. We assume that the responses are polytomous (i.e., we have more than two options per question), i.e., $Y_{ij} \in \{1, \dots, M_j\}$, where $M_j > 2$ denotes the number of options for question j.

Note that the most basic form of the IRT model [19], [25] is dichotomous, modeling binary-valued (i.e., graded as correct/incorrect) responses, which can be seen as a special case of the general polytomous model with $M_j = 2$. We allow the number of options M_j to vary across different questions. In many practical scenarios, not every response Y_{ij} is observed. Consequently, let Ω_{obs} denote the index set of observed entries in the data matrix $\mathbf{Y}$.

We assume that a predictor variable Z_{ij} induces a probability distribution over the set of M_j options for student i to select on question j. There are many models available for defining the predictor Z_{ij}, including linear regression [5], [13], low-rank

TABLE 2
Mean and standard deviation of the prediction error rate of SPRITE, (L)ORD [16], NRM [6], and GPCM [21] on various datasets over 50 random splits of the data set. SPRITE achieves the best prediction performance on every dataset.

Dataset description	SPRITE	(L)ORD	NRM	GPCM
Algebra test	**0.25 (0.01)**	0.29 (0.02)	0.31 (0.01)	0.26 (0.01)
Computer engineering course	**0.17 (0.01)**	0.31 (0.02)	0.33 (0.01)	0.21 (0.01)
Probability course	**0.41 (0.01)**	0.68 (0.03)	0.57 (0.01)	0.62 (0.01)
Signals and systems course	**0.29 (0.01)**	0.48 (0.02)	0.41 (0.01)	0.56 (0.01)
Comprehensive university exam	**0.53 (0.01)**	0.71 (0.01)	0.63 (0.01)	0.62 (0.01)

models [18], [26], [37], and cluster-based models [8]. Without loss of generality, we will restrict ourselves to the IRT models with uni-dimensional student ability parameters and no guessing parameters, i.e., the 1-PL and 2-PL IRT models. We will also not consider multi-dimensional IRT (MIRT) models with multi-dimensional student ability parameters, e.g., [11], [27]. Concretely, the dichotomous 1PL IRT model defines a latent parameter $\theta_i \in \mathbb{R}$ for each student $i = 1, \ldots, N$, as well as a latent parameter $\alpha_j \in \mathbb{R}$ for each question $j = 1, \ldots, Q$[2]. The predictor variable Z_{ij} is then given by $Z_{ij} = \theta_i - \alpha_j$. In an educational context, θ_i corresponds to ith student's ability and α_j corresponds to jth question's intrinsic difficulty. 2PL IRT models further introduce a discrimination parameter β_j for each question that characterizes the ability question j can separate students with high abilities from students with low abilities. The predictor variable Z_{ij} is then given by $Z_{ij} = \beta_j(\theta_i - \alpha_j)$.

2.2 Existing categorical IRT models

There exist a number of extensions to the basic IRT model that are suitable for questions with categorical options, e.g., the nominal response model (NRM) [6] based on the softmax function [15]. The NRM uses independent exponential functions to model each categorical option. For the NRM, the probability that student i will select option y for question j is defined as

$$P(Y_{ij} = y|\theta_i, \boldsymbol{\beta_j}, \boldsymbol{\alpha_j}) = \frac{\exp\big(\beta_{jy}(\theta_i - \alpha_{jy})\big)}{\sum_{k=1}^{M_j} \exp\big(\beta_{jk}(\theta_i - \alpha_{jk})\big)},$$

where θ_i is the latent trait of student i, $\boldsymbol{\beta_j} = [\beta_{j1}, \ldots, \beta_{jM_j}]^T$ is a vector of discrimination factors that characterizes the capability of the options in the j^{th} question in discriminating students with different latent traits, and $\boldsymbol{\alpha_j} = [\alpha_{j1}, \ldots, \alpha_{jM_j}]^T$ is a vector of offset parameters for the options in the j^{th} question. The NRM does not explicitly model the order of the options nor does it take into account which option is the correct option in each question to constrain its item parameters. As a result, its latent trait parameter does not necessarily correlate with student abilities. Therefore, the NRM is more suited to psychological tests rather than educational scenarios, where it is crucial to estimate student abilities from data.

2.3 Existing ordinal IRT models

We now detail existing ordinal IRT models, including the generalized partial credit model (GPCM) [21] and the ordinal response model (ORD) [16], [17] that are suitable for ordinal options.

2. Most IRT models assume that the ability parameters are randomly drawn from a given distribution and therefore only estimate the question parameters and not the student parameters.

2.3.1 GPCM

The GPCM [21] is a generalized version of the strictly ordinal partial credit model [20] that attempts to allow partial ordering of the options. It is closely related to the graded response model [28], both widely adapted in real-world testing applications. The GPCM is constructed from successive dichotomization of adjacent options. Under the GPCM, the probability that student i will select option y for question j is defined as

$$P(Y_{ij} = y|\theta_i, \beta_j, \boldsymbol{\alpha_j}) = \frac{\exp\big(\sum_{v=1}^{y} \beta_j(\theta_i - \alpha_{jv})\big)}{\sum_{k=1}^{M_j} \exp\big(\sum_{v=1}^{k} \beta_j(\theta_i - \alpha_{jv})\big)}, \quad (1)$$

where β_j is a single, non-negative discrimination factor for the jth question and the vector $\boldsymbol{\alpha_j} = [\alpha_{j1}, \ldots, \alpha_{jM_j}]^T$ represents threshold values where adjacent options have equal probability of being chosen. We refer the reader to [21] for the derivation of (1). Intuitively, the GPCM says that the probability of selecting option y is proportional to the probability of successively comparing a set of adjacent pairs of options (i.e., for a student to select Option C, they have to first select Option B over Option A, and then select Option C over Option B, and finally select Option C over Option D). Intuitively, this construction of successive dichotomous decisions among the options of a question in the GPCM model still tries to enforce ordinal options, although its option parameters are not necessarily strictly ordered. Technically, the GPCM still requires a known ordering of options as input, and is best applicable to test the validity of a given ordering rather than to learn a partial order directly from observed data.

2.3.2 ORD

The standard ordinal response (ORD) model [16], [17], [38] is the most commonly used model for ordinal data in machine learning literature, assuming a known and fixed ordering of the options. The ORD model posits a latent ability variable $Z'_{ij}, \forall (i,j) \in \Omega_{\text{obs}}$, defined as

$$Z'_{ij} = Z_{ij} + \varepsilon_{ij} = \theta_i - \alpha_j + \varepsilon_{ij}, \quad (2)$$

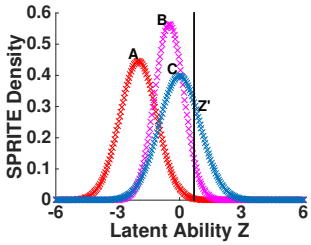
where ε_{ij} is a standard normal random variable. The model further imposes a set of ordered bin positions on the jth question denoted by $-\infty = \gamma_j^0 < \gamma_j^1 < \cdots < \gamma_j^{M_j} = \infty$, which map the latent predictor variable Z'_{ij} into one of the M_j options as follows

$$Y_{ij} = y \quad \text{if} \quad \gamma_j^{y-1} < Z'_{ij} \leq \gamma_j^y, \quad y \in \{1, \ldots, M_j\}.$$

A common constraint imposed on the bin positions is $\gamma_j^1 = 0$, which avoids identifiability problems where the bin positions could be shifted and scaled without changing the likelihood of the observed data [16].

Using the definition of the variable Z'_{ij} in (2), the probability of selecting option $y \in \{1, \ldots, M_j\}$ is given by

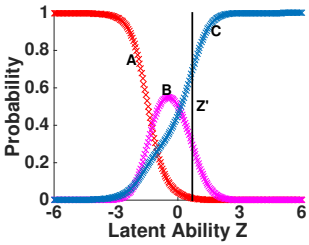
$$P(Y_{ij} = y|Z_{ij}) = \Phi(\gamma_j^y - Z_{ij}) - \Phi(\gamma_j^{y-1} - Z_{ij}).$$



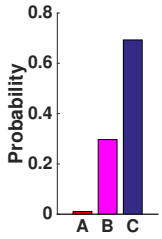
(a) A set of three Gaussian option functions or so-called "sprites."

$$p(Y = y|Z) = \frac{\mathcal{N}(Z|\mu^y, \nu^y)}{\sum_{k=1}^{M} \mathcal{N}(Z|\mu^k, \nu^k)}$$

(b) The probabilistic model underlying SPRITE.



(c) The probabilities of selecting each option under the SPRITE model.



(d) Probabilities of selecting each option for latent ability Z' in (a).

Fig. 1. Illustration of the SPRITE model for a multiple-choice question with one correct option and two incorrect options (a total of $M = 3$ options) that are incorrect to different degrees; The flexibility of SPRITE enables us to model strictly ordered as well as partially ordered options. The location of the latent ability variable Z and the Gaussian option functions (referred to as "sprites") induce a probability mass function determining the probability of the option y (out of A, B, and C) a student will select. (a) The sprites associated with each option. We set the mean and variance parameters of the correct option (in this case, C) to $\mu^k = 0$ and $\nu^k = 1$. The other incorrect options have $\mu^k \leq 0$. (b) The SPRITE probability model. (c) The resulting choice probabilities as a function of Z (known as item category response functions (ICRFs) in IRT literature). (d) The probabilities of selecting each option for a particular latent-predictor value of Z', indicated by the vertical line in (a) and (c).

Here, $\Phi(x)$ denotes the cumulative distribution function of the standard normal distribution, which is defined as $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp(-\frac{t^2}{2}) \mathrm{d}t$. The ORD model is only applicable when a strict option ordering is given a priori.

3 The SPRITE Model

We now introduce the SPRITE model. The ORD model discussed in Section 2.3.2 combines the latent predictor Z_{ij} with the bin positions contained in $\boldsymbol{\gamma}_j$ in order to generate the observed response Y_{ij}. The SPRITE model, by contrast, does not rely on a set of bins, but rather on a set of distributions over the latent space of student ability that each option corresponds to (see Figure 1 for an illustration of this principle). For each question j, each option $k \in \{1, \ldots, M_j\}$ specifies a Gaussian function with mean μ_j^k and variance ν_j^k. We call each option's Gaussian function a *sprite*. We model the probability that student i will select option y of question j given the value of the latent predictor variable Z_{ij} as follows

$$P(Y_{ij} = y|Z_{ij}) = \frac{\mathcal{N}(Z_{ij}|\mu_j^y, \nu_j^y)}{\sum_{k=1}^{M_j} \mathcal{N}(Z_{ij}|\mu_j^k, \nu_j^k)},$$

where $\mathcal{N}(Z|\mu, \nu)$ denotes a Gaussian distribution with mean μ and variance ν. As described in Section 2.1, $Z_{ij} = \theta_i - \alpha_j$. Since one can arbitrarily shift the values of the question difficulty parameter α_j and also shift the values of the option mean parameters μ_j^k accordingly without changing the option selection likelihood, we will omit the α_j parameters and only keep the option mean parameters μ_j^k. Thus, the latent predictor variable Z_{ij} is simply given by $Z_{ij} = Z_i$. We therefore re-write the likelihood equation associated with SPRITE using only the student latent ability parameters Z_i as follows:

$$P(Y_{ij} = y|Z_i) = \frac{\mathcal{N}(Z_i|\mu_j^y, \nu_j^y)}{\sum_{k=1}^{M_j} \mathcal{N}(Z_i|\mu_j^k, \nu_j^k)}. \quad (3)$$

Figure 1 (a) shows the item option density functions or "sprites" of each of the three options induced over the space of the latent student ability parameter Z. Figure 1 (c) shows the item category response functions (ICRFs) that characterizes the probability of selecting each option as a function of Z. We emphasize that the model setup of SPRITE enables the flexibility of partially ordered options since the sprites of each option can overlap.

Like all polytomous IRT models, the SPRITE model parameters can be susceptible to identifiability issues. For example, one can negate all the learned option means μ_j^k and at the same time negate the inferred student latent ability parameters Z_i without affecting the model likelihood. To prevent such identifiability issues, we fix the mean of the sprite that corresponds to the correct option to zero and its variance to one. We further constrain the sprites that correspond to the incorrect options to have non-positive means. These restrictions enable us to interpret the latent ability parameters of each student, as larger values of Z_i means that a student is more likely to select the correct option. Moreover, the option mean parameters are also interpretable as smaller values of μ_j^k of an option indicate a higher degree of incorrectness, in the sense that this option is more likely to be selected by students having low latent ability.

We note that in applications where there is no information on which option is correct for a given question, we can simply fix the sprite of an arbitrary option to have zero mean and unit variance.

4 Parameter Inference for SPRITE

We now detail our inference algorithm for the SPRITE model. We first note that, under the Bayesian setting, there exist a number of methods for fitting SPRITE to data. We will rely on Markov Chain Monte-Carlo sampling methods [13], which are easy to deploy for our model. Unlike expectation maximization (EM) [5], [12], which produces point estimates, an MCMC-based approach provides full posterior distributions on each parameter of interest.

4.1 MCMC sampler for SPRITE

The prior distributions on each of the latent parameter of interest are assumed to be

$$\mu_j^k \sim \mathcal{N}^-(\mu_\mu, \nu_\mu), \;\; Z_i \sim \mathcal{N}(\mu_z, \nu_z), \;\; \nu_j^k \sim \mathcal{IG}(\alpha_\nu, \beta_\nu), \quad (4)$$

where $\mathcal{IG}(\alpha, \beta)$ denotes the inverse gamma distribution with shape parameter α and scale parameter β, and $\mathcal{N}^-(\mu, \nu)$ denotes the truncated normal distribution on the region $(-\infty, 0]$ with mean μ and variance ν. μ_μ, ν_μ, α_ν, β_ν, μ_z, ν_z are hyperparameters for the prior distributions of the latent option mean, option variance, and student latent ability parameters.

We present a Metropolis-within-Gibbs sampler [14] for SPRITE. The SPRITE latent variables $\mathbf{Z}$, $\boldsymbol{\mu}_j$, and $\boldsymbol{\nu}_j$ for

$i = 1, \ldots, N$ and $j = 1, \ldots, Q$ are sampled via a Metropolis–Hastings step at each MCMC iteration. Here, we introduce the vector notation $\boldsymbol{\mu}_j = [\mu_j^1, \ldots, \mu_j^{M_j}]^T$, $\boldsymbol{\nu}_j = [\nu_j^1, \ldots, \nu_j^{M_j}]^T$, and $\mathbf{Z} = [Z_1, \ldots, Z_N]^T$. Furthermore, we treat the missing observations in $\mathbf{Y}$ as latent variables and sample them using Gibbs sampling. A summary of the steps used by our MCMC sampler is as follows. We use the notation $[\cdot]^t$ to represent the state of a parameter at iteration t, for $t = 1, \ldots, T$ where T denotes the total number of MCMC iterations.

1) Propose new latent abilities using a normal random walk, i.e., $[Z_i]^t \sim \mathcal{N}\left([Z_i]^{t-1}, \sigma_z^2\right)$ for $i = 1, \ldots, N$.
2) Propose new option means using a truncated normal random walk, i.e., $[\mu_j^k]^t \sim \mathcal{N}^-\left([\mu_j^k]^{t-1}, \sigma_\mu^2\right)$ for $j = 1, \ldots, Q$ and $k = 1, \ldots, M_j$.
3) Propose new option variances using a log-normal random walk, i.e., $\log[\nu_j^k]^t \sim \mathcal{N}(\log[\nu_j^k]^{t-1}, \sigma_\nu^2)$.
4) Calculate a Metropolis-Hastings acceptance/rejection probability based on the likelihood ratio r between proposed parameters and the parameters from the previous MCMC iteration and the proposal transition probabilities. The proposed latent variables $[Z_i]^t$, $[\mu_j^k]^t$, and $[\nu_j^k]^t$ for $i = 1, \ldots, N$, $j = 1, \ldots, Q$ and $k = 1, \ldots, M_j$ are then jointly accepted or rejected.
5) Propose new prediction values for the missing responses $[Y_{ij}]^t$, $(i, j) \notin \Omega_{\text{obs}}$ by sampling the probabilities induced by (3) using $[Z_i]^t$, $[\mu_j^k]^t$, and $[\nu_j^k]^t$.

In the above, σ_z^2, σ_μ^2, and σ_ν^2 are user-defined tuning parameters that control the variance of the random walk proposal distributions. The likelihood ratio r for the Metropolis-Hastings acceptance/rejection step is given by

$$\prod_{i,j} \frac{p([Y_{i,j}]^{t-1} | [Z_i]^t, [\boldsymbol{\mu}_j]^t, [\boldsymbol{\nu}_j]^t)}{p([Y_{i,j}]^{t-1} | [Z_i]^{t-1}, [\boldsymbol{\mu}_j]^{t-1}, [\boldsymbol{\nu}_j]^{t-1})}$$
$$\times \prod_i \frac{p([Z_i]^t | \mu_z, \nu_z)}{p([Z_i]^{t-1} | \mu_z, \nu_z)} \prod_j \frac{p([\boldsymbol{\mu}_j]^t | \mu_\mu, \nu_\mu) p([\boldsymbol{\nu}_j]^t | \alpha_\nu, \beta_\nu)}{p([\boldsymbol{\mu}_j]^{t-1} | \mu_\mu, \nu_\mu) p([\boldsymbol{\nu}_j]^{t-1} | \alpha_\nu, \beta_\nu)}$$
$$\times \prod_{i,j,k} \frac{\mathcal{N}([\mu_j^k]^{t-1} | [\mu_j^k]^t, \sigma_\mu^2) \Phi(-\frac{[\mu_j^k]^{t-1}}{\sigma_\mu^2})}{\mathcal{N}([\mu_j^k]^t | [\mu_j^k]^{t-1}, \sigma_\mu^2) \Phi(-\frac{[\mu_j^k]^t}{\sigma_\mu^2})},$$

where the data likelihood terms are given by (3) and the prior likelihood terms are given by (4). The last term corresponds to the asymmetric proposal transition probabilities for the option means.

4.2 Posterior inference

After a suitable burn-in period, the MCMC sampler from Section 4.1 produces samples that approximate the true posterior distribution of all model parameters. We will make use of the posterior means of the parameters Z_i, μ_j^k, and ν_j^k when performing experiments in which we compare the estimated model parameters to a known ground truth. For real data experiments, we predict unobserved option selections using the posterior mode of $Y_{ij}, (i, j) \notin \Omega_{\text{obs}}$.

5 Experiments

We first evaluate SPRITE using synthetic data to demonstrate model convergence, identifiability, and consistency. We then compare the predictive performance of SPRITE to other IRT models detailed in Section 2 using real-world educational datasets.

5.1 Synthetic data experiments

We first generate the ground truth SPRITE model parameters Z_i, $\boldsymbol{\mu}_j$, and $\boldsymbol{\nu}_j$ for $i = 1, \ldots, N$ and $j = 1, \ldots, Q$. For simplicity of exposition, we fix the number of options per question to $M_j = M = 5$, $\forall j$. We also set the number of questions to equal the number of students, i.e., $Q = N \in \{50, 100, 150\}$. We generate the latent parameters via (4) and the options each student select on each question $\mathbf{Y}$ via (3). In this experiment, the response matrix $\mathbf{Y}$ is fully observed. The hyperparameters are as follows: $\mu_z = 0$, $\nu_z = 1$, $\nu_\mu = 1$, $\alpha_\nu = 1$, and $\beta_\nu = 1$.

We deploy SPRITE as described in Section 4.1 by initializing all parameters of interest with random values. We use 90,000 iterations in the burn-in phase of our MCMC sampler and compute the posterior means for all parameters as described in Section 4.2 over an additional 10,000 iterations. We compare the estimated SPRITE model parameters to their known ground truth values using the following three error metrics

$$E_z = \frac{\|\hat{\mathbf{Z}} - \mathbf{Z}\|_2^2}{\|\mathbf{Z}\|_2^2}, \quad E_\mu = \frac{\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2^2}{\|\boldsymbol{\mu}\|_2^2}, \quad E_\nu = \frac{\|\hat{\boldsymbol{\nu}} - \boldsymbol{\nu}\|_2^2}{\|\boldsymbol{\nu}\|_2^2}, \quad (5)$$

where $\boldsymbol{\mu} = [\boldsymbol{\mu}_1^T, \ldots, \boldsymbol{\mu}_Q^T]^T$, $\boldsymbol{\nu} = [\boldsymbol{\nu}_1^T, \ldots, \boldsymbol{\nu}_Q^T]^T$, and $\| \cdot \|_2$ denotes the ℓ_2-norm. The quantities $\hat{\mathbf{Z}}$, $\hat{\boldsymbol{\mu}}$, and $\hat{\boldsymbol{\nu}}$ represent model estimated values computed as described in Section 4.1 and $\mathbf{Z}$, $\boldsymbol{\mu}$, and $\boldsymbol{\nu}$ represent the known ground-truth values.

Figure 2 shows box-whisker plots for the three error metrics in (5) for various problem sizes. The low error rates and the convergence of the parameter estimates to their ground truth values demonstrate that the MCMC sampling algorithm is capable of estimating the SPRITE model parameters accurately. Furthermore, all error metrics decrease as the problem size increases, which implies model consistency.

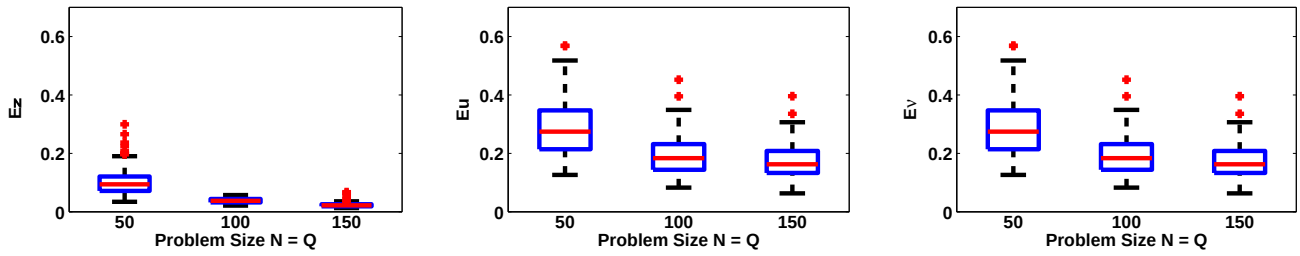
5.2 Real-world data experiments

In this section, we demonstrate the superiority of SPRITE over other polytomous IRT models using a variety of real-world datasets. We first demonstrate that SPRITE outperforms other models in terms of prediction accuracy on unobserved student responses, and then showcase the interpretability of the SPRITE model parameters.

5.2.1 Predictive performance

We now compare the predictive performance of SPRITE on unobserved student responses against the NRM, GPCM, and the ORD model (described in Section 2). In order to test the ORD model on datasets with unknown question option orderings, we slightly modify the MCMC parameter inference algorithm of the ORD model proposed in [16]. This is done by additionally sampling a permutation of option orderings at each step of the MCMC. For M options, there are $M!$ such permutations. We simply propose a new option ordering sampled uniformly from the set of all possible permutations and incorporate this additional proposal step into the final Metropolis-Hastings rejection ratio. We refer to this modified algorithm as the *learned ordinal (LORD) algorithm*.

We study five educational datasets consisting of students' responses to multiple-choice questions. A brief description of the datasets can be found in Table 3. The "algebra test" dataset is from a secondary level algebra test administered on Amazon's Mechanical Turk [18]. The datasets "computer engineering course," "probability course," and "signals and systems course" are from college level courses administered on OpenStax Tutor [23]. Each of these datasets contain a number of missing entires—corresponding



(a) Student latent ability error E_z. (b) Option mean error E_μ. (c) Option variance error E_ν.

Fig. 2. Synthetic experiment over various problem sizes (number of students N and number of questions Q) where $N = Q$, and $M_j = M = 5$ options. (a) The error of the latent ability E_z; (b) The error of the option means E_μ; (c) The error of option variance E_ν. All three error metrics decrease as the problem size (the number of students and questions) grows.

to the case where students did not answer every question. Finally, the "comprehensive university exam" dataset contains responses on a university level comprehensive exam [35]. The number of options M_j of each question varies for every dataset except the comprehensive university exam dataset, where each question has $M_j = 4$ options.

The NRM analyzes every dataset as if the options in each question are categorical. For the algebra test dataset, a domain expert has provided an ordering to the options as a reference (according to the incorrectness of each option). In this case, we use this expert ordering as the input to both the GPCM and the ORD model. We fix the sprite for the correct option to have zero mean and unit variance on each question for the SPRITE model. On the other datasets, we do not have any information on the ordering of the options in each question or which option is correct. In this case, for the GPCM, we use an arbitrary ordering of the options as input. We further use the LORD variant of the standard ORD model for these datasets to provide the best possible comparison. We set the sprite for an arbitrary option to have zero mean and unit variance, and remove the non-positive constraint on the means of the sprites of other options for the SPRITE model.

We compare the predictive performance of the algorithms by first puncturing the full dataset, i.e., removing a portion of the entries in the observed student response matrix $\mathbf{Y}$ uniformly at random and use these values as a test set. We set the rate of puncturing to be 20%. We then train each model using the remaining observed data and make predictions on the test set as discussed in Section 4.2. The error metric is simply the ratio of the number of incorrect predictions over the total number of predictions made. All experiments were repeated over 50 random puncturing patterns of the dataset. We use 90,000 MCMC sampling iterations for the burn-in period and compute posterior statistics over an additional 10,000 iterations. The hyperparameters are set as follows: $\mu_z = -1$, $\nu_z = 1$, $\nu_\mu = 1$, $\alpha_\nu = 1$, and $\beta_\nu = 1$. The random walk variance parameters are set as $\sigma_z^2 = \sigma_\mu^2 = \sigma_\nu^2 = 0.01$. In our experiments, we have found that the performance of the MCMC sampling algorithm is robust against the values of these parameters.

The predictive performance results are summarized in Table 2. SPRITE outperforms all other models on all datasets, sometimes by a significant margin. This performance gain is likely due to the fact that the SPRITE model learns an ordering directly from data (unlike the GPCM), and the fact that it allows the options to be partially ordered (unlike the NRM and the ORD model).

TABLE 4
An example question and its options from the algebra test dataset.

Option	If $\frac{5x}{3} + 1 = -9$, then $x =$?
A	-6
B	$\frac{50}{3}$
C	$-\frac{24}{5}$
D	$-\frac{50}{3}$
E	$\frac{30}{5}$

5.2.2 *Interpreting the SPRITE model parameters*

One major advantage of the SPRITE model lies in the interpretability of its model parameters. In addition, our model allows the options to be partially ordered, and we can learn this ordering solely from data without any effort from human experts. We now examine the option orderings estimated by the SPRITE model on the algebra test dataset, and compare it against the ordering provided by the human expert. An example question in this dataset is listed in Table 4.

The estimated parameters for each option is listed in Table 5. Option A is the correct option, while Option C is the least incorrect option, as students who selected it have mastered the key concept of fraction manipulations, but made an error of not negating the $+1$ on the left hand side of the equation when moving it to the right hand side. Options D and B are estimated to be the almost equally incorrect and more incorrect than other options, as students who selected them have not mastered fraction manipulations. Option E lies between Option C and Options D and B, but with its huge estimated variance, it overlaps with every other option. This option ordering learned from data matches the expert ordering except for Option E, which is specified to be equally incorrect as Option C. It is interesting to see the large variance estimated for Option E as it can be considered as an "outlier" option (it is the only option that is not fully reduced). Therefore, it can be identified as a very uninformative option as it does not provide much information towards the latent abilities of students that select it (its SPRITE density has little variation across a large range of Z_i values). Therefore, Option E can possibly be removed from this question in favor of other more informative distractor options.

This illustrative example shows that SPRITE is capable of learning a (partial) ordering of the options automatically from student response data. Moreover, these findings can often provide interesting insights into each option, and potentially, coupled with

TABLE 3
Description of datasets. Unobserved data listed in the table refers to actual missing responses in the respective datasets; Q denotes the number of questions in each dataset and N denotes the number of students.

Description	Size ($Q \times N$)	Maximum no. options	Known ordering	Observed data
Algebra test	34×99	5	Yes	100%
Computer engineering course	203×82	12	No	97%
Probability course	86×49	7	No	67%
Signals and systems course	143×44	11	No	64%
Comprehensive university exam	60×1567	4	No	71%

TABLE 5
Estimated SPRITE parameters for each option for the question from the algebra test dataset shown in Table 4.

Option	A	B	C	D	E
Mean	0.00	−0.43	−0.21	−0.44	−0.30
Variance	1.00	0.41	0.55	0.40	3.64

expert knowledge, improve the quality of the questions.

5.2.3 A notion of question informativeness

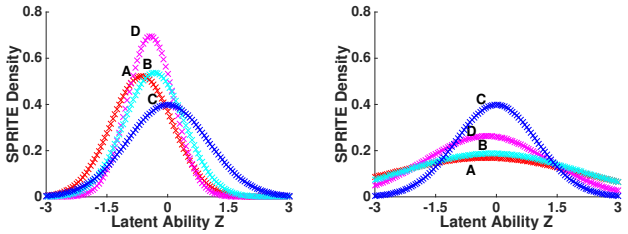
As demonstrated above, the parameters $\boldsymbol{\mu}_j$ and $\boldsymbol{\nu}_j$ of the SPRITE model provide an intuitive ordering of the options in a question and information on the informativeness of each option. We now extend this observation to a notion of the informativeness of each question, which is crucial in educational scenarios as the role of assessment questions is to use the students' responses to estimate the latent abilities of the students Z_i. This notion of the informativeness of a question on the latent ability of a student will help us to compare the quality of individual questions.

Using the SPRITE model, one direct notion of question informativeness is the mutual information (MI), denoted by $I(Z;Y_j)$ and measured in bits, between a students' latent ability $Z_i \in \mathbb{R}$ and the options they select $Y_{i,j} \in \{1,\dots,M_j\}$ on a question. A question with large MI is more informative than a question with low MI. The MI is formally defined as

$$\begin{aligned} I(Z;Y_j) &= \sum_{y=1}^{M_j} \int P(Y=y,Z) \log_2 \frac{P(Y=y,Z)}{P(Z)P(Y=y)} \mathrm{d}Z \\ &= \sum_{y=1}^{M_j} \int P(Y=y|Z) \mathrm{P}(Z) \log_2 \frac{P(Y=y|Z)}{P(Y=y)} \mathrm{d}Z. \quad (6) \end{aligned}$$

Here, $P(Y=y|Z)$ is the SPRITE model likelihood given by (3), $P(Z) = \mathcal{N}(Z|\mu_z,\nu_z)$ is the Gaussian prior on latent abilities given by (4), and $\mathrm{P}(Y=y) = \int P(Y=y|Z)P(Z)\mathrm{d}Z$ is the marginal distribution of selecting each option. The integral in (6) is difficult to evaluate in closed-form. However, using student ability estimates $\hat{Z}_i$, $\forall i$ and option mean and variance estimates $\hat{\boldsymbol{\mu}}_j$, $\hat{\boldsymbol{\nu}}_j$ it can be easily approximated numerically for each question.

Figure 3 shows one informative and one uninformative question in the algebra test dataset. The informative question (MI = 0.38 bits) illustrated in Figure 3(a) reveals that the sprite that corresponds to the correct option dominates in the region $Z > 0$, meaning that it is likely to be selected for students with high latent ability. The other options are more likely to be selected by students with low latent abilities. This means that on this question, which option a student selects is very informative about their latent ability. By contrast, the less informative question (MI = 0.03 bits) illustrated in Figure 3(b) reveals multiple overlapping options that show little discriminative power. In other words, which option a student selects on this question does not provide much information about their latent ability.



(a) An informative question with MI of 0.38 bits (b) An uninformative question with MI of 0.03 bits

Fig. 3. Estimated informativeness of two questions from the algebra test dataset. The curves represent Gaussian option functions, or sprites.

We emphasize that this notion of question informativeness is particularly useful in computerized adaptive testing (CAT) [32], [33], [35], where one wants to select questions that are the most informative in estimating a student's ability out of a large collection of questions.

6 Conclusion

In this paper, we have developed the SPRITE (stochastic polytomous response item) model to analyze students' responses to multiple-choice questions, which is capable of automatically learning a (partial) ordering among the options of each question solely from data. SPRITE outperforms existing, state-of-the-art polytomous IRT models in terms of predicting unobserved student responses on five real-world educational datasets. Additionally, we have demonstrated that the estimated SPRITE model parameters provide interesting insights regarding the ordering of the options on each question that can be used to improve the quality of the questions. Moreover, our results have shown that SPRITE also provides a natural notion of informativeness of each question that can be used to select the most informative questions out of a large collection of questions in adaptive testing applications.

Several future directions look promising. First, improvements to the MCMC sampler could potentially improve the computational efficiency of the SPRITE parameter inference algorithm. Methods such as variational Bayes [1], expectation maximization [5], and Metropolis–Hastings Robbins–Monro [9] may sacrifice little in terms of data fitting performance while providing significant savings in computation time. Additionally, alternative models for the predictor variable Z, such as MIRT [4] and linear regression models with either fixed or learned covariates [18], [24], could provide

additional improvements in terms of prediction performance and interpretability.

ACKNOWLEDGMENTS

Thanks to the Chairman, JAC, IISER Pune, for sharing the "comprehensive university exam" dataset, as well as Divyanshu Vats for insightful discussion regarding this dataset. This work was partially supported by the grants ARO W911NF–15–1–0316 and AFOSR FA9550–14–1–0088.

REFERENCES

[1] H. Attias. Inferring parameters and structure of latent variable models by variational Bayes. In *Proc. 15th Conf. on Uncertainty in Artificial Intelligence*, pages 21–30, July 1999.

[2] T. Barnes. The Q-matrix method: Mining student response data for knowledge. In *Proc. AAAI Workshop Educational Data Mining*, Pittsburg, PA, July 2005.

[3] T. M. Bechger, G. Maris, H. Verstralen, and A. A Béguin. Using classical test theory in combination with item response theory. *Applied Psychological Measurement*, 27(5):319–334, Sep. 2003.

[4] A. A. Béguin and C. A. Glas. MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika*, 66(4):541–561, Dec. 2001.

[5] C. M. Bishop and N. M. Nasrabadi. *Pattern Recognition and Machine Learning*. Springer New York, 2006.

[6] R. D. Bock. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37(1):29–51, Mar. 1972.

[7] P. Brusilovsky and C. Peylo. Adaptive and intelligent web-based educational systems. *Intl. J. Artificial Intelligence in Education*, 13(2-4):159–172, Apr. 2003.

[8] L. M. Busse, P. Orbanz, and J. M. Buhmann. Cluster analysis of heterogeneous rank data. In *Proc. 24th Intl. Conf. on Machine learning*, pages 113–120, June 2007.

[9] L Cai. Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *J. Educational and Behavioral Statistics*, 35(3):307–335, June 2010.

[10] A. T. Corbett and J. R. Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-adapted Interaction*, 4(4):253–278, Dec. 1994.

[11] J. De La Torre. A cognitive diagnosis model for cognitively based multiple-choice options. *Applied Psychological Measurement*, 33(3):163–183, May 2009.

[12] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[13] A. Gelman, C. Robert, N. Chopin, and J. Rousseau. *Bayesian Data Analysis*. CRCPress, 1995.

[14] W. R. Gilks, N. G. Best, and K. K. Tan. Adaptive rejection Metropolis sampling sithin Gibbs sampling. *Applied Statistics*, 44(4):455–472, 1995.

[15] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2010.

[16] V. E. Johnson and J. H. Albert. *Ordinal Data Modeling*. Springer New York, 1999.

[17] A. S. Lan, C. Studer, A. E. Waters, and R. G. Baraniuk. Tag-aware ordinal sparse factor analysis for learning and content analytics. In *Proc. 6th Intl. Conf. Educational Data Mining*, pages 90–97, July 2013.

[18] A. S. Lan, A. E. Waters, C. Studer, and R. G. Baraniuk. Sparse factor analysis for learning and content analytics. *J. Machine Learning Research*, 15:1959–2008, June 2014.

[19] F. M. Lord. *Applications of Item Response Theory to Practical Testing Problems*. Erlbaum Associates, 1980.

[20] G. N. Masters. A Rasch model for partial credit scoring. *Psychometrika*, 47(2):149–174, June 1982.

[21] E. Muraki. A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16(2):159–176, June 1992.

[22] M. L. Nering and R. Ostini. *Handbook of Polytomous Item Response Theory Models*. Taylor & Francis, 2011.

[23] OpenStax Tutor. URL http://openstaxtutor.org/, 2014.

[24] P. I. Pavlik, H. Cen, and K. R. Koedinger. Learning factors transfer analysis: Using learning curve analysis to automatically generate domain models. In *Proc. 2nd Intl. Conf. on Educational Data Mining*, pages 121–130, July 2009.

[25] G. Rasch. *Probabilistic Models for Some Intelligence and Attainment Tests*. MESA Press, 1993.

[26] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, Aug. 2010.

[27] M. D. Reckase. *Multidimensional Item Response Theory*. Springer Publishing Company, Incorporated, 2009.

[28] F. Samejima. Graded response model. In *Handbook of Modern Item Response Theory*, pages 85–100. 1997.

[29] K. K. Tatsuoka. A probabilistic model for diagnosing misconceptions by the pattern classification approach. *J. Educational and Behavioral Statistics*, 10(1):55–73, Mar. 1985.

[30] D. Thissen and L. Steinberg. *A Response Model for Multiple-choice Items*. Springer, 1997.

[31] D. Thissen, L. Steinberg, and A. R. Fitzpatrick. Multiple-choice models: The distractors are also part of the item. *J. Educational Measurement*, 26(2):161–176, June 1989.

[32] W. J. van der Linden. Bayesian item selection criteria for adaptive testing. *Psychometrika*, 63(2):201–216, June 1998.

[33] W. J. van der Linden and C. Glas. *Computerized Adaptive Testing: Theory and Practice*. Kluwer Academic Publishers, 2000.

[34] K. VanLehn, C. Lynch, K. Schulze, J. A. Shapiro, R. Shelby, L. Taylor, D. Treacy, A. Weinstein, and M. Wintersgill. The Andes physics tutoring system: Lessons learned. *Intl. J. Artificial Intelligence in Education*, 15(3):147–204, Aug. 2005.

[35] D. Vats, C. Studer, A. S. Lan, L. Carin, and R. G. Baraniuk. Test size reduction for concept estimation. In *Proc. 6th Intl. Conf. on Educational Data Mining*, pages 292–295, July 2013.

[36] J. E. Ware, J. B. Bjorner, and M. Kosinski. Practical implications of item response theory and computerized adaptive testing: A brief summary of ongoing studies of widely used headache impact scales. *Medical Care*, 38(9):II73–II82, Sep. 2000.

[37] M. Zhou, C. Wang, M. Chen, J. Paisley, D. Dunson, and L. Carin. Nonparametric Bayesian matrix completion. In *IEEE Sensor Array and Multichannel Signal Processing Workshop*, pages 213–216, Oct. 2010.

[38] A. Zymnis, S. Boyd, and E. Candès. Compressed sensing with quantized measurements. *IEEE Sig. Proc. Letters*, 17(2):149–152, Feb. 2010.