# Variational Item Response Theory: Fast, Accurate, and Expressive

Mike Wu [1], Richard L. Davis [2], Benjamin W. Domingue [2], Chris Piech [1], Noah Goodman [1,3]

Department of Computer Science [1], Education [2], and Psychology [3]

Stanford University

{wumike,rldavis,bdomingu,cpiech,ngoodman}@stanford.edu

## ABSTRACT

Item Response Theory (IRT) is a ubiquitous model for understanding humans based on their responses to questions, used in fields as diverse as education, medicine and psychology. Large modern datasets offer opportunities to capture more nuances in human behavior, potentially improving test scoring and better informing public policy. Yet larger datasets pose a difficult speed / accuracy challenge to contemporary algorithms for fitting IRT models. We introduce a variational Bayesian inference algorithm for IRT, and show that it is fast and scaleable without sacrificing accuracy. Using this inference approach we then extend classic IRT with expressive Bayesian models of responses. Applying this method to five large-scale item response datasets from cognitive science and education yields higher log likelihoods and improvements in imputing missing data. The algorithm implementation is open-source, and easily usable.

## 1. INTRODUCTION

The task of estimating human ability from stochastic responses to a series of questions has been studied since the 1950s in thousands of papers spanning several fields. The standard statistical model for this problem, Item Response Theory (IRT), is used every day around the world, in many critical contexts including college admissions tests, school-system assessment, survey analysis, popular questionnaires, and medical diagnosis.

As datasets become larger, new challenges and opportunities for improving IRT models present themselves. On the one hand, massive datasets offer the opportunity to better understand human behavior, fitting more expressive models. On the other hand, the algorithms that work for fitting small datasets often become intractable for larger data sizes. Indeed, despite a large body of literature, contemporary IRT methods fall short – it remains surprisingly difficult to estimate human ability from stochastic responses. One crucial bottleneck is that the most accurate, state-of-the-art Bayesian inference algorithms are prohibitively slow, while

faster algorithms (such as the popular maximum marginal likelihood estimators) are less accurate and poorly capture uncertainty. This leaves practitioners with a choice: either have nuanced Bayesian models with appropriate inference or have timely computation.

In the field of artificial intelligence, a revolution in deep generative models via *variational inference* [25, 37] has demonstrated an impressive ability to perform fast inference for complex Bayesian models. In this paper, we present a novel application of variational inference to IRT, validate the resulting algorithms with synthetic datasets, and apply them to real world datasets. We then show that this inference approach allows us to extend classic IRT response models with deep neural network components. We find that these more flexible models better fit the large real world datasets. Specifically, our contributions are as follows:

1. **Variational inference for IRT:** We derive a new optimization objective — the Variational Item response theory Lower Bound, or VIBO — to perform inference in IRT models. By learning a mapping from responses to posterior distributions over ability and items, VIBO is "amortized" to solve inference queries efficiently.

2. **Faster inference:** We find VIBO to be much faster than previous Bayesian techniques and usable on much larger datasets without loss in accuracy.

3. **More expressive:** Our inference approach is naturally compatible with deep generative models and, as such, we enable the novel extension of Bayesian IRT models to use neural-network-based representations for inputs, predictions, and student ability. We develop the first deep generative IRT models.

4. **Simple code:** Using our VIBO python package[1] is only a few lines of code that is easy to extend.

5. **Real world application:** We demonstrate the impact of faster inference and expressive models by applying our algorithms to datasets including: PISA, DuoLingo and Gradescope. We achieve up to 200 times speedup and show improved accuracy at imputing hidden responses. At scale, these improvements in efficiency save hundreds of hours of computation.

---

[1] http://github.com/mhw32/variational-item-response-theory-public

As a roadmap, in Sec. 2 we describe the item response theory challenge. In Sec. 3 we present a main algorithm. Finally, in Sec. 4 and 5 we show its impact on speed and accuracy.

## 2. BACKGROUND

We briefly review several variations of item response theory and the fundamental principles of approximate Bayesian inference, focusing on modern variational inference.

### 2.1 Item Response Theory Review

Imagine answering a series of multiple choice questions. For example, consider a personality survey, a homework assignment, or a school entrance examination. Selecting a response to each question is an interaction between your "ability" (knowledge or features) and the characteristics of the question, such as its difficulty. The goal in examination analysis is to gauge this unknown ability of each student and the unknown item characteristics based only on responses. Early procedures [11] defaulted to very simple methods, such as counting the number of correct responses, which ignore differences in question quality. In reality, we understand that not all questions are created equal: some may be hard to understand while others may test more difficult concepts. To capture these nuances, Item Response Theory (IRT) was developed as a mathematical framework to reason jointly about people's ability and the items themselves.

The IRT model plays an impactful role in many large institutions. It is the preferred method for estimating ability in several state assessments in the United States, for international assessments gauging educational competency across countries [18], and for the National Assessment of Educational Programs (NAEP), a large-scale measurement of literacy and reading comprehension in the US [35]. Beyond education, IRT is a method widely used in cognitive science and psychology, for instance with regards to studies of language acquisition and development [19, 28, 14, 7].
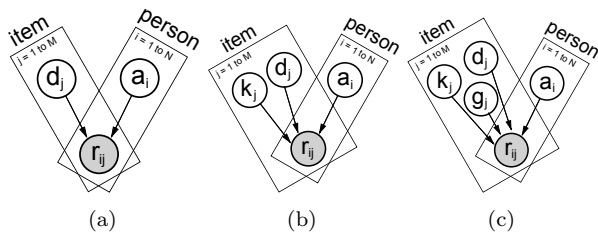


Figure 1: Graphical models for the (a) 1PL, (b) 2PL, and (c) 3PL Item Response Theories. Observed variables are shaded. Arrows represent dependency between random variables and each rectangle represents a plate (i.e. repeated observations).

IRT has many forms; we review the most standard (Fig. 1). The simplest class of IRT summarizes the ability of a person with a single parameter. This class contains three versions: 1PL, 2PL, and 3PL IRT, each of which differ by the number of free variables used to characterize an item. The 1PL IRT model, also called the Rasch model [34], is given in Eq. 1,

$$p(r_{i,j} = 1 | a_i, d_j) = \frac{1}{1 + e^{-(a_i - d_j)}} \quad (1)$$

where $r_{i,j}$ is the response by the $i$-th person to the $j$-th item. There are $N$ people and $M$ items in total. Each item in the 1PL model is characterized by a single number representing difficulty, $d_j$. As the 1PL model is equivalent to a logistic function, a higher difficulty requires a higher ability in order to respond correctly. Next, the 2PL IRT model[2] adds a *discrimination* parameter, $k_j$ for each item that controls the slope (or scale) of the logistic curve. We can expect items with higher discrimination to more quickly separate people of low and high ability. The 3PL IRT model further adds a *pseudo-guessing* parameter, $g_j$ for each item that sets the asymptotic minimum of the logistic curve. We can interpret pseudo-guessing as the probability of success if the respondent were to make a reasonable guess on an item. The 2PL and 3PL IRT models are:

$$p(r_{i,j} | a_i, \boldsymbol{d}_j) = \frac{1}{1 + e^{-k_j a_i - d_j}} \text{ or } g_j + \frac{1 - g_j}{1 + e^{-k_j a_i - d_j}} \quad (2)$$

where $\boldsymbol{d}_j = \{k_j, d_j\}$ for 2PL and $\boldsymbol{d}_j = \{g_j, k_j, d_j\}$ for 3PL. See Fig. 1 for graphical models of each of these IRT models.

A single ability dimension is sometimes insufficient to capture the relevant variation in human responses. For instance, if we are measuring a person's understanding on elementary arithmetic, then a single dimension may suffice in capturing the majority of the variance. However, if we are instead measuring a person's general mathematics ability, a single real number no longer seems sufficient. Even if we bound the domain to middle school mathematics, there are several factors that contribute to "mathematical understanding" (e.g. proficiency in algebra versus geometry). Summarizing a person with a single number in this setting would result in a fairly loose approximation. For cases where multiple facets of ability contribute to performance, we consider *multidimensional* item response theory [1, 36, 29]. We focus on 2PL multidimensional IRT (MIRT):

$$p(r_{i,j} = 1 | \boldsymbol{a}_i, \boldsymbol{k}_j, d_j) = \frac{1}{1 + e^{-\mathbf{a}_i^T \boldsymbol{k}_j - d_j}} \quad (3)$$

where we use bolded notation $\boldsymbol{a}_i = (a_i^{(1)}, a_i^{(2)}, \dots a_i^{(K)})$ to represent a $K$ dimensional vector. Notice that the item discrimination becomes a vector of equal size to ability.

In practice, given a (possibly incomplete) $N \times M$ matrix of observed responses, we want to *infer* the ability of all $N$ people and the characteristics of all $M$ items. Next, we provide a brief overview of inference in IRT.

### 2.2 Inference in Item Response Theory

Inference is the task of estimating unknown variables, such as ability, given observations, such as student responses. We compare and contrast three popular methods used to perform inference for IRT in research and industry. Inference algorithms are critical for item response theory as slow or inaccurate algorithms prevent the use of appropriate models.

**Maximum Likelihood Estimation** A straightforward approach is to pick the most likely ability and item features

---

[2]We default to 2PL as the pseudo-guessing parameter introduces several invariances in the model. This requires far more data to infer ability accurately, as measured by our own synthetic experiments. For practitioners, we warn against using 3PL for small to medium datasets.

given the observed responses. To do so we optimize:

$$\mathcal{L}_{\text{MLE}} = \max_{\{\boldsymbol{a}_i\}_{i=1}^N, \{\boldsymbol{d}_j\}_{j=1}^M} \sum_{i=1}^N \sum_{j=1}^M \log p(r_{ij}|\boldsymbol{a}_i, \boldsymbol{d}_j) \quad (4)$$

with stochastic gradient descent (SGD). The symbol $\boldsymbol{d}_j$ represents all item features e.g. $\boldsymbol{d}_j = \{d_j, \boldsymbol{k}_j\}$ for 2PL. Eq. 4 is often called the Joint Maximum Likelihood Estimator [3, 12], abbreviated MLE. MLE poses inference as a supervised regression problem in which we choose the most likely unknown variables to match known dependent variables. While MLE is simple to understand and implement, it lacks any measure of uncertainty; this can have important consequences especially when responses are missing.

**Expectation Maximization** Several papers have pointed out that when using MLE, the number of unknown parameters increases with the number of people [5, 17]. In particular, [17] shows that in practical settings with a finite number of items, standard convergence theorems do not hold for MLE as the number of people grows. To remedy this, the authors instead treat ability as a nuisance parameter and marginalized it out [5, 6]. Brock et. al. introduces an Expectation-Maximization (EM) [10] algorithm to iterate between (1) updating beliefs about item characteristics and (2) using the updated beliefs to define a marginal distribution (without ability) $p(r_{ij}|\boldsymbol{d}_j)$ by numerical integration of $\boldsymbol{a}_i$. Appropriately, this algorithm is referred to as Maximum Marginal Likelihood Estimation, which we abbreviate as EM. Eq. 6 shows the E and M steps for EM.

$$\text{E step}: \quad p(r_{ij}|\boldsymbol{d}_j^{(t)}) = \int_{\boldsymbol{a}_i} p(r_{ij}|\boldsymbol{a}_i, \boldsymbol{d}_j^{(t)}) p(\boldsymbol{a}_i) d\boldsymbol{a}_i \quad (5)$$

$$\text{M step}: \quad \boldsymbol{d}_j^{(t+1)} = \arg\max_{\boldsymbol{d}_j} \sum_{i=1}^N \log p(r_{ij}|\boldsymbol{d}_j^{(t)}) \quad (6)$$

where $(t)$ represents the iteration count. We often choose $p(\boldsymbol{a}_i)$ to be a simple prior distribution like standard Normal. In general, the integral in the E-step is intractable: EM uses a Gaussian-Hermite quadrature to discretely approximate $p(r_{ij}|\boldsymbol{d}_j^{(t)})$. See [20] for a closed form expression for $\boldsymbol{d}_j^{(t+1)}$ in the M step. This method finds the maximum a posteriori (MAP) estimate for item characteristics. EM does not infer ability as it is "ignored" in the model: the common workaround is to use EM to infer item characteristics, then fit ability using a second auxiliary model. In practice, EM has grown to be ubiquitous in industry as it is incredibly fast for small to medium sized datasets. However, we expect that EM may scale poorly to large datasets and higher dimensions as numerical integration requires far more points to properly measure a high dimensional volume.

**Hamiltonian Monte Carlo** The two inference methods above give only point estimates for ability and item characteristics. In contrast Bayesian approaches seek to capture the full posterior over ability and item characteristics given observed responses, $p(\boldsymbol{a}_i, \boldsymbol{d}_{1:M}|\boldsymbol{r}_{i,1:M})$ where $\boldsymbol{r}_{i,1:M} = (r_{i,1}, \cdots, r_{i,M})$. Doing so provides estimates of uncertainty and characterizes features of the joint distribution that cannot be represented by point estimates, such as multimodality and parameter correlation. In practice, this can be very useful for a more robust understanding of student ability.

The common technique for Bayesian estimation in IRT uses Markov Chain Monte Carlo (MCMC) [21, 15] to draw samples from the posterior by constructing a Markov chain carefully designed such that $p(\boldsymbol{a}_i, \boldsymbol{d}_{1:M}|\boldsymbol{r}_{i,1:M})$ is the equilibrium distribution. By running the chain longer, we can closely match the distribution of drawn samples to the true posterior. Hamiltonian Monte Carlo (HMC) [33, 32, 23] is an efficient version of MCMC for continuous state spaces. We recommend [23] for a good review of HMC.

The strength of this approach is that the samples generated capture the true posterior (if the algorithm is run long enough). Bute the computational costs for MCMC can be very high, and the cost scales at least linearly with the number of latent parameters — which for IRT is proportional to data size. With new datasets of millions of observations, such limitations can be debilitating. Fortunately, there exist a second class of approximate Bayesian techniques that have gained significant traction in the machine learning community. We provide a careful review of *variational inference*.

## 2.3 Variational Methods Review

Variational inference (VI) first appeared from the statistical physics community and was later generalized for many probabilistic models by Jordan et. al. [24]. In recent years, VI has been popularized in machine learning where it is used to do inference in large graphical models decribing images and natural language. The main intuition of variational inference is to treat inference as an optimization problem: starting with a family of distributions, the goal is to pick the one that best approximates the true posterior, by minimizing an estimate of the mismatch between true and approximate distributions. We will first describe VI in the general context of a latent variable model, and then apply VI to IRT.

Let $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{z} \in \mathcal{Z}$ represent observed and latent variables, respectively. (In the context of IRT, $\boldsymbol{x}$ represents the responses from a single student and $\boldsymbol{z}$ represents ability and item characteristics.) In VI [24, 42, 4], we introduce a family of tractable distributions over $\boldsymbol{z}$ such that we can easily sample from and score. We wish to find the member $q_{\psi^*(\boldsymbol{x})} \in \mathcal{Q}$ that minimizes the Kullback-Leibler (KL) divergence between itself and the true posterior:

$$q_{\psi^*(\boldsymbol{x})}(\boldsymbol{z}) = \arg\min_{q_{\psi(\boldsymbol{x})}} D_{\text{KL}}(q_{\psi(\boldsymbol{x})}(\boldsymbol{z})||p(\boldsymbol{z}|\boldsymbol{x})) \quad (7)$$

where $\psi(\boldsymbol{x})$ are parameters that define each distribution. For example, $\psi(\boldsymbol{x})$ would be the mean and scale for a Gaussian distribution. Since the "best" approximate posterior $q_{\psi^*(\boldsymbol{x})}$ depends on the observed variables, its parameters have $\boldsymbol{x}$ as a dependent variable. To be clear, there is one approximate posterior for every possible value of the observed variables.

Frequently, we need to do inference for many different values of $\boldsymbol{x}$. For example, student A and student B may have picked different answers to the same question. Since their responses differ, we would need to do inference twice. Let $p_{\mathcal{D}}(\boldsymbol{x})$ be an empirical distribution over the observed variables, which is equivalent to the marginal $p(\boldsymbol{x})$ if the generative model is correctly specified. Then, the average quality of the variational approximations is measured by

$$\mathbb{E}_{p_{\mathcal{D}}(\boldsymbol{x})} \left[ \max_{\psi(\boldsymbol{x})} \mathbb{E}_{q_{\psi(\boldsymbol{x})}(\boldsymbol{z})} \left[ \log \frac{p(\boldsymbol{x}, \boldsymbol{z})}{q_{\psi(\boldsymbol{x})}(\boldsymbol{z})} \right] \right] \quad (8)$$

In practice, $p_\mathcal{D}(\boldsymbol{x})$ is unknown but we assume access to a dataset $\mathcal{D}$ of examples i.i.d. sampled from $p_\mathcal{D}(\boldsymbol{x})$; this is sufficient to evaluate Eq. 8.

**Amortization** As in Eq. 8, we must learn an approximate posterior for each $\boldsymbol{x} \in \mathcal{D}$. For a large dataset $\mathcal{D}$, this can quickly grow to be unwieldly. One such solution to this scalability problem is *amortization* [16], which reframes the per-observation optimization problem as a supervised regression task. Consider learning a single deterministic mapping $f_\phi : \mathcal{X} \to \mathcal{Q}$ to predict $\psi^*(\boldsymbol{x})$ or equivalently $q_{\psi^*(\boldsymbol{x})} \in \mathcal{Q}$ as a function of the observation $\boldsymbol{x}$. Often, we choose $f_\phi$ to be a conditional distribution, denoted by $q_\phi(\boldsymbol{z}|\boldsymbol{x}) = f_\phi(\boldsymbol{x})(\boldsymbol{z})$.

The benefit of amortization is a large reduction in computational cost: the number of parameters is vastly smaller than learning a per-observation posterior. Additionally, if we manage to learn a good regressor, then the amortized approximate posterior $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ could generalize to new observations $\boldsymbol{x} \notin \mathcal{D}$ unseen in training. This strength has made amortized VI popular with modern latent variable models, such as the Variational Autoencoder [25].

Instead of Eq. 8, we now optimize:

$$\max_\phi \mathbb{E}_{p_\mathcal{D}(\boldsymbol{x})} \left[ \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})} \left[ \log \frac{p(\boldsymbol{x},\boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x})} \right] \right] \qquad (9)$$

The drawback of this approach is that it introduces an *amortization gap*: since we are technically using a less flexible family of approximate distributions, the quality of approximate posteriors can be inferior.

**Model Learning** So far we have assumed a fixed generative model $p(\boldsymbol{x}, \boldsymbol{z})$. However, often we can only specify a family of possible models $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ parameterized by $\theta$. The symmetric challenge (to approximate inference) is to choose $\theta$ whose model best explains the evidence. Naturally, we do so by maximizing the log marginal likelihood of the data

$$\log p_\theta(\boldsymbol{x}) = \log \int_{\boldsymbol{z}} p_\theta(\boldsymbol{x}, \boldsymbol{z}) d\boldsymbol{z} \qquad (10)$$

Using Eq. 9, we derive the Evidence Lower Bound (ELBO) [25, 37] with $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ as our inference model

$$\log p_\theta(\boldsymbol{x}) \geq \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})} \left[ \log \frac{p_\theta(\boldsymbol{x},\boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x})} \right] \triangleq \text{ELBO} \qquad (11)$$

We can jointly optimize $\phi$ and $\theta$ to maximize the ELBO. We have the option to parameterize $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ and $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ with deep neural networks, as is common with the VAE [25], yielding an extremely flexible space of distributions.

**Stochastic Gradient Estimation** The gradients of the ELBO (Eq. 11) with respect to $\phi$ and $\theta$ are:

$$\nabla_\theta \text{ELBO} = \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\nabla_\theta \log p_\theta(\boldsymbol{x}, \boldsymbol{z})]] \qquad (12)$$

$$\nabla_\phi \text{ELBO} = \nabla_\phi \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}, \boldsymbol{z})] \qquad (13)$$

Eq. 12 can be estimated using Monte Carlo samples. However, as it stands, Eq. 13 is difficult to estimate as we cannot distribute the gradient inside the inner expectation. For certain families $\mathcal{Q}$, we can use a reparameterization trick.

**Reparameterization Estimators** Reparameterization is the technique of removing sampling from the gradient

computation graph [25, 37]. In particular, if we can reduce sampling $\boldsymbol{z} \sim q_\phi(\boldsymbol{z}|\boldsymbol{x})$ to sampling from a parameter-free distribution $\epsilon \sim p(\epsilon)$ plus a deterministic function application, $\boldsymbol{z} = g_\phi(\epsilon)$, then we may rewrite Eq. 13 as:

$$\nabla_\phi \text{ELBO} = \mathbb{E}_{p(\epsilon)}[\nabla_{\boldsymbol{z}} \log \frac{p_\theta(\boldsymbol{x}, \boldsymbol{z}(\epsilon))}{q_\phi(\boldsymbol{z}(\epsilon)|\boldsymbol{s})} \nabla_\phi g_\phi(\epsilon)] \qquad (14)$$

which now can be estimated efficiently by Monte Carlo (the gradient is inside the expectation). A benefit of reparameterization over alternative estimators (e.g. score estimator [30] or REINFORCE [44]) is lower variance while remaining unbiased. A common example is if $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ is Gaussian $\mathcal{N}(\mu, \sigma^2)$ and we choose $p(\epsilon)$ to be $\mathcal{N}(0, 1)$, then $g(\epsilon) = \epsilon * \sigma + \mu$.

## 3. THE VIBO ALGORITHM

Having rehearsed the major principles of VI, we will adapt them to IRT. In our review, we presented the ELBO that serves as the primary loss function to train an inference model. Given the nuances of IRT, we can derive a new loss function specialized for ability and item characteristics. We call the resulting algorithm VIBO since it is a **V**ariational approach for **I**tem response theory based on a novel lower **BO**und. While the remainder of the section presents the technical details, we ask the reader to keep the high-level purpose in mind: VIBO is an objective function that if we maximize, we have a method to predict student ability from his or her responses. As a optimization problem, VIBO is much cheaper computationally than MCMC.

To show that doing so is justifiable, we prove that VIBO well-defined. That is, we must show that VIBO lower bounds the marginal likelihood over a student's responses.

THEOREM 3.1. *Let $\boldsymbol{a}_i$ be the ability for person $i \in [1, N]$ and $\boldsymbol{d}_j$ be the characteristics for item $j \in [1, M]$. We use the shorthand notation $\boldsymbol{d}_{1:M} = (\boldsymbol{d}_1, \ldots, \boldsymbol{d}_M)$. Let $r_{i,j}$ be the binary response for item $j$ by person $i$. We write $\boldsymbol{r}_{i,1:M} = (r_{i,1}, \ldots r_{i,M})$. If we define the VIBO objective as:*

$$\text{VIBO} \triangleq \mathcal{L}_{recon} + \mathbb{E}_{q_\phi(\boldsymbol{d}_{1:M}|\boldsymbol{r}_{i,1:M})}[D_{ability}] + D_{item}$$

*where*

$$\mathcal{L}_{recon} = \mathbb{E}_{q_\phi(\boldsymbol{a}_i, \boldsymbol{d}_{1:M}|\boldsymbol{r}_{i,1:M})}[\log p_\theta(\boldsymbol{r}_{i,1:M}|\boldsymbol{a}_i, \boldsymbol{d}_{1:M})]$$
$$D_{ability} = D_{\text{KL}}(q_\phi(\boldsymbol{a}_i|\boldsymbol{d}_{1:M}, \boldsymbol{r}_{i,1:M})||p(\boldsymbol{a}_i))$$
$$D_{item} = D_{\text{KL}}(q_\phi(\boldsymbol{d}_{1:M}|\boldsymbol{r}_{i,1:M})||p(\boldsymbol{d}_{1:M}))$$

*and assume the joint posterior factors as $q_\phi(\boldsymbol{a}_i, \boldsymbol{d}_{1:M}|\boldsymbol{r}_{i,1:M}) = q_\phi(\boldsymbol{a}_i|\boldsymbol{d}_{1:M}, \boldsymbol{r}_{i,1:M})q_\phi(\boldsymbol{d}_{1:M}|\boldsymbol{r}_{i,1:M})$, then $\log p(\boldsymbol{r}_{i,1:M}) \geq \text{VIBO}$. In othe words, VIBO is a lower bound on the log marginal probability of person $i$'s responses.*

PROOF. Expand marginal and apply Jensen's inequality:

$$\log p_\theta(\boldsymbol{r}_{i,1:M}) \geq \mathbb{E}_{q_\phi(\boldsymbol{a}_i, \boldsymbol{d}_{1:M}|\boldsymbol{r}_{i,1:M})} \left[ \log \frac{p_\theta(\boldsymbol{r}_{i,1:M}, \boldsymbol{a}_i, \boldsymbol{d}_{1:M})}{q_\phi(\boldsymbol{a}_i, \boldsymbol{d}_{1:M}|\boldsymbol{r}_{i,1:M})} \right]$$
$$= \mathbb{E}_{q_\phi(\boldsymbol{a}_i, \boldsymbol{d}_{1:M}|\boldsymbol{r}_{i,1:M})}[\log p_\theta(\boldsymbol{r}_{i,1:M}|\boldsymbol{a}_i, \boldsymbol{d}_{1:M})]$$
$$+ \mathbb{E}_{q_\phi(\boldsymbol{a}_i, \boldsymbol{d}_{1:M}|\boldsymbol{r}_{i,1:M})} \left[ \log \frac{p(\boldsymbol{a}_i)}{q_\phi(\boldsymbol{a}_i|\boldsymbol{d}_{1:M}, \boldsymbol{r}_{i,1:M})} \right]$$
$$+ \mathbb{E}_{q_\phi(\boldsymbol{a}_i, \boldsymbol{d}_{1:M}|\boldsymbol{r}_{i,1:M})} \left[ \log \frac{p(\boldsymbol{d}_{1:M})}{q_\phi(\boldsymbol{d}_{1:M}|\boldsymbol{r}_{i,1:M})} \right]$$
$$= \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{A}} + \mathcal{L}_{\text{B}}$$

Rearranging the latter two terms, we find that:

$$\mathcal{L}_A = \mathbb{E}_{q_\phi(\boldsymbol{d}_{1:M}|\boldsymbol{r}_{i,1:M})}\left[D_{\mathrm{KL}}(q_\phi(\boldsymbol{a}_i|\boldsymbol{d}_{1:M},\boldsymbol{r}_{i,1:M})||p(\boldsymbol{a}_i))\right]$$

$$\mathcal{L}_B = \mathbb{E}_{q_\phi(\boldsymbol{d}_{1:M}|\boldsymbol{r}_{i,1:M})}\left[\log\frac{p(\boldsymbol{d}_{1:M})}{q_\phi(\boldsymbol{d}_{1:M}|\boldsymbol{r}_{i,1:M})}\right]$$

$$= D_{\mathrm{KL}}(q_\phi(\boldsymbol{d}_{1:M}|\boldsymbol{r}_{i,1:M})||p(\boldsymbol{d}_{1:M}))$$

Since $\mathrm{VIBO} = \mathcal{L}_{\mathrm{recon}} + \mathcal{L}_A + \mathcal{L}_B$, and KL terms are non-negative, we have shown that VIBO bounds $\log p_\theta(\boldsymbol{r}_{i,1:M})$. $\square$

Thm. 3.1 leaves several choices up to us, and we opt for the simplest ones. For instance, the prior distributions are chosen to be independent standard Normal distributions: $p(\boldsymbol{a}_i) = \prod_{k=1}^{K} p(a_{i,k})$ and $p(\boldsymbol{d}_{1:M}) = \prod_{j=1}^{M} p(\boldsymbol{d}_j)$ where $p(a_{i,k})$ and $p(\boldsymbol{d}_j)$ are $\mathcal{N}(0,1)$. Further, we found it sufficient to assume $q_\phi(\boldsymbol{d}_{1:M}|\boldsymbol{r}_{i,1:M}) = q_\phi(\boldsymbol{d}_{1:M}) = \prod_{j=1}^{M} q_\phi(\boldsymbol{d}_j)$ although nothing prevents the general case. Initially, we assume the generative model, $p_\theta(\boldsymbol{r}_{i,1:M}|\boldsymbol{a}_i,\boldsymbol{d}_{1:M})$, to be an IRT model (thus $\theta$ is empty); later we explore generalizations.

---

**Algorithm 1:** VIBO Forward Pass

---
Assume we are given observed responses for person $i$, $\boldsymbol{r}_{i1:M}$;
Compute $\mu_{\boldsymbol{d}}, \sigma_{\boldsymbol{d}}^2 = q_\phi(\boldsymbol{d}_{1:M})$;
Sample $\boldsymbol{d}_{1:M} \sim \mathcal{N}(\mu_{\boldsymbol{d}}, \sigma_{\boldsymbol{d}}^2)$;
Compute $\mu_{\boldsymbol{a}}, \sigma_{\boldsymbol{a}}^2 = q_\phi(\boldsymbol{a}_i|\boldsymbol{d}_{1:M},\boldsymbol{r}_{i,1:M})$;
Sample $\boldsymbol{a}_i \sim \mathcal{N}(\mu_{\boldsymbol{a}}, \sigma_{\boldsymbol{a}}^2)$;
Compute $\mathcal{L}_{\mathrm{recon}} = \log p_\theta(\boldsymbol{r}_{i,1:M}|\boldsymbol{a}_i,\boldsymbol{d}_{1:M})$;
Compute $D_{\mathrm{ability}} = D_{\mathrm{KL}}(\mathcal{N}(\mu_{\boldsymbol{a}}, \sigma_{\boldsymbol{a}}^2)||\mathcal{N}(0,1))$;
Compute $D_{\mathrm{item}} = D_{\mathrm{KL}}(\mathcal{N}(\mu_{\boldsymbol{d}}, \sigma_{\boldsymbol{d}}^2)||\mathcal{N}(0,1))$;
Compute $\mathrm{VIBO} = \mathcal{L}_{\mathrm{recon}} + D_{\mathrm{ability}} + D_{\mathrm{item}}$

---

The posterior $q_\phi(\boldsymbol{a}_i|\boldsymbol{d}_{1:M},\boldsymbol{r}_{i,1:M})$ needs to be robust to missing data as often not every person answers every question. To achieve this, we explore the following family:

$$q_\phi(\boldsymbol{a}_i|\boldsymbol{d}_{1:M},\boldsymbol{r}_{i,1:M}) = \prod_{j=1}^{M} q_\phi(\boldsymbol{a}_i|\boldsymbol{d}_j,\boldsymbol{r}_{i,j}) \qquad (15)$$

If we assume each component $q_\phi(\boldsymbol{a}_i|\boldsymbol{d}_j,\boldsymbol{r}_{i,j})$ is Gaussian, then $q_\phi(\boldsymbol{a}_i|\boldsymbol{d}_{1:M},\boldsymbol{r}_{i,1:M})$ is Gaussian as well, being a Product-Of-Experts [22, 45]. If item $j$ is missing, we replace its term in the product with the prior, $p(\boldsymbol{a}_i)$ representing no added information. We found this design to outperform averaging over non-missing entries: $\frac{1}{M}\sum_{j=1}^{M} q_\phi(\boldsymbol{a}_i|\boldsymbol{d}_j,\boldsymbol{r}_{i,j})$.

As VIBO is a close cousin of the ELBO, we can estimate its gradients with respect to $\theta$ and $\phi$ similarly:

$$\nabla_\theta \mathrm{VIBO} = \nabla_\theta \mathcal{L}_{\mathrm{recon}}$$
$$= \mathbb{E}_{q_\phi(\boldsymbol{a}_i,\boldsymbol{d}_{1:M}|\boldsymbol{r}_{i,1:M})}\left[\nabla_\theta \log p_\theta(\boldsymbol{r}_{i,1:M}|\boldsymbol{a}_i,\boldsymbol{d}_{1:M})\right]$$
$$\nabla_\phi \mathrm{VIBO} = \nabla_\phi \mathbb{E}_{q_\phi(\boldsymbol{d}_{1:M}|\boldsymbol{r}_{i,1:M})}[D_{\mathrm{ability}}] + \nabla_\phi D_{\mathrm{item}}$$
$$= \nabla_\phi \mathbb{E}_{q_\phi(\boldsymbol{a}_i,\boldsymbol{d}_{1:M}|\boldsymbol{r}_{i,1:M})}\left[\frac{p(\boldsymbol{a}_i)p(\boldsymbol{d}_{1:M})}{q_\phi(\boldsymbol{a}_i,\boldsymbol{d}_{1:M}|\boldsymbol{r}_{i,1:M})}\right]$$

As in Eq. 14, we may wish to move the gradient inside the KL divergences by reparameterization to reduce variance. To allow easy reparameterization, we define all variational distributions $q_\phi(\cdot|\cdot)$ as Normal distributions with diagonal covariance. In practice, we find that estimating $\nabla_\theta \mathrm{VIBO}$ and $\nabla_\phi \mathrm{VIBO}$ with a single sample is sufficient. With this setup,

VIBO can be optimized using stochastic gradient descent to learn an amortized inference model that maximizes the marginal probability of observed data. We summarize the required computation to calculate VIBO in Alg. 1.

## 4. DATASETS

We explore one synthetic dataset, to build intuition and confirm parameter recovery, and five large scale applications of IRT to real world data, summarized in Table 1.

Table 1: Dataset Statistics

|  | # Persons | # Items | Missing Data? |
|---|---|---|---|
| CritLangAcq | 669498 | 95 | N |
| WordBank | 5520 | 797 | N |
| DuoLingo | 2587 | 2125 | Y |
| Gradescope | 1254 | 98 | Y |
| PISA | 519334 | 183 | Y |

**Synthetic IRT** To sanity check that VIBO performs as well as other inference techniques, we synthetically generate a dataset of responses using a 2PL IRT model: sample $\boldsymbol{a}_i \sim p(\boldsymbol{a}_i)$, $\boldsymbol{d}_j \sim p(\boldsymbol{d}_j)$. Given ability and item characteristics, IRT-2PL determines a Bernoulli distribution over responses to item $j$ by person $i$. We sample once from this Bernoulli distribution to "generate" an observation. In this setting, we know the ground truth ability and item characteristics. We vary $N$ and $M$ to explore parameter recovery.

**Second Language Acquisition** This dataset contains native and non-native English speakers answering questions to a grammar quiz[3], which upon completion would return a prediction of the user's native language. Using social media, over half a million users of varying ages and demographics completed the quiz. Quiz questions often contain both visual and linguistic components. For instance, a quiz question could ask the user to "choose the image where the dog is chased by the cat" and present two images of animals where only one of image agrees with the caption. Every response is thus binary, marked as correct or incorrect. In total, there are 669,498 people with 95 items and no missing data. The creators of this dataset use it to study the presence or absence of a "critical period" for second language acquisition [19]. We will refer to this dataset as CritLangAcq.

**WordBank: Vocabulary Development** The MacArthur-Bates Communicative Development Inventories (CDIs) are a widely used metric for early language acquisition in children, testing concepts in vocabulary comprehension, production, gestures, and grammar. The WordBank [14] database archives many independently collected CDI datasets across languages and research laboratories[4]. The database consists of a matrix of people against vocabulary words where the $(i,j)$ entry is 1 if a parent reports that child $i$ has knowledge of word $j$ and 0 otherwise. Some entries are missing due to slight variations in surveys and incomplete responses. In total, there are 5,520 children responding to 797 items.

---

[3] The quiz can be found at www.gameswithwords.org. The data is publically available at osf.io/pyb8s.
[4] github.com/langcog/wordbankr

**DuoLingo: App-Based Language Learning** We examine the 2018 DuoLingo Shared Task on Second Language Acquisition Modeling[5] [38]. This dataset contains anonymized user data from the popular education application, DuoLingo. In the application, users must choose the correct vocabulary word among a list of distractor words. We focus on the subset of native English speakers learning Spanish and only consider lesson sessions. Each user has a timeseries of responses to a list of vocabulary words, each of which is shown several times. We repurpose this dataset for IRT: the goal being to infer the user's language proficiency from his or her errors. As such, we average over all times a user has seen each vocabulary item. For example, if the user was presented "habla" 10 times and correctly identified the word 5 times, he or she would be given a response score of 0.5. We then round to 0 or 1. We revisit a continuous version Sec. 7. After processing, we have 2587 users and 2125 vocabulary words with missing data as users frequently drop out. We ignore user and syntax features.

**Gradescope: Course Exam Data** Gradescope [39] is a course application that assists teachers in grading student assignments. This dataset contains 105,218 reponses from 6,607 assignments in 2,748 courses and 139 schools. All assignments are instructor-uploaded, fixed-template assignments, with at least 3 questions, with the majority being examinations. We focus on course 102576, randomly chosen. We remove students who did not respond to any questions and round up partial credit. In total, there are 1254 students with 98 items, with missing entries.

**PISA 2015: International Assessment** The Programme for International Student Assessment (PISA) is an international exam that measures 15-year-old students' reading, mathematics, and science literacy every three years. It is run by the Organization for Economic Cooperation and Development (OECD). The OECD released anonymized data from PISA '15 for students from 80 countries and education systems[6]. We focus on the science component. Using IRT to access student performance is part of the pipeline the OECD uses to compute holistic literacy scores for different countries. As part of our processing, we binarize responses, rounding any partial credit to 1. In total, there are 519,334 students and 183 questions. Not every student answers every question as many versions of the computer exam exist.

## 5. FAST AND ACCURATE INFERENCE

We will show that VIBO is as accurate as HMC and nearly as fast as MLE/EM, making Bayesian IRT a realistic, even preferred, option for modern applications.

### 5.1 Evaluation

We compare compute cost of VIBO to HMC, EM[7], and MLE using IRT-2PL by measuring wall-clock run time. For HMC, we limit to drawing 200 samples with 100 warmup steps with no parallelization. For VIBO and MLE, we use the Adam optimizer with a learning rate of 5e-3. We choose to conservatively optimize for 10k iterations to estimate cost.

---

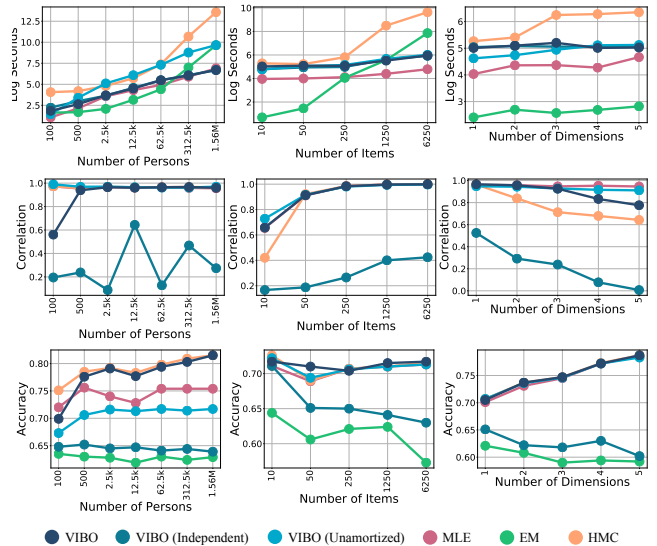[7] We use the popular MIRT package in R for EM with 61 points for numerical integration.



Figure 2: Performance of inference algorithms for IRT for synthetic data, as we vary the number of people, items, and latent ability dimensions. (Top) Computational cost in log-seconds (e.g. 1 log second is about 3 seconds whereas 10 log seconds is 6.1 hours). (Middle) Correlation of inferred ability with true ability (used to generate the data). (Bottom) Accuracy of held-out data imputation.

However, speed only matters assuming good performance. We use three metrics of accuracy: (1) For the synthetic dataset, because we know the true ability, we can measure the expected correlation between it and the inferred ability under each algorithm (with the exception of EM as ability is not inferred). A correlation of 1.0 would indicate perfect inference. (2) The most general metric is the accuracy of imputed missing data. We hold out 10% of the responses, use the inferred ability and item characteristics to generate responses thereby populating missing entries, and compute prediction accuracy for held-out responses. This metric is a good test of "overfitting" to observed responses. (3) In the case of fully Bayesian methods (HMC and VIBO) we can compare posterior predictive statistics [40] to further test uncertainty calibration (which accuracy alone does not capture). Recall that the posterior predictive is defined as:

$$p(\tilde{\boldsymbol{r}}_{i,1:M}|\boldsymbol{r}_{i,1:M}) = \mathbb{E}_{p(\boldsymbol{a}_i,\boldsymbol{d}_{1:M}|\boldsymbol{r}_{i,1:M})}[p(\tilde{\boldsymbol{r}}_{i,1:M}|\boldsymbol{a}_i,\boldsymbol{d}_{1:M})]$$

For HMC, we have samples of ability and item characteristics from the true posterior whereas for VIBO, we draw samples from the $q_\phi(\boldsymbol{a}_i,\boldsymbol{d}_{1:M}|\boldsymbol{r}_{i,1:M})$. Given such parameter samples, we can then sample responses. We compare summary statistics of these response samples: the average number of items answered correctly per person and the average number of people who answered each item correctly.

### 5.2 Synthetic Data Results

With synthetic experiments we are free to vary $N$ and $M$ to extremes to stress test the inference algorithms: first, we range from 100 to 1.5 million people, fixing the number of items to 100 with dimensionality 1; second, we range from 10 to 6k items, fixing 10k people with dimesionality 1; third, we vary the number of latent ability dimensions from 1 to 5, keeping a constant 10k people and 100 items.

(a) Real World: Accuracy of Imputing Missing Data vs Time Cost



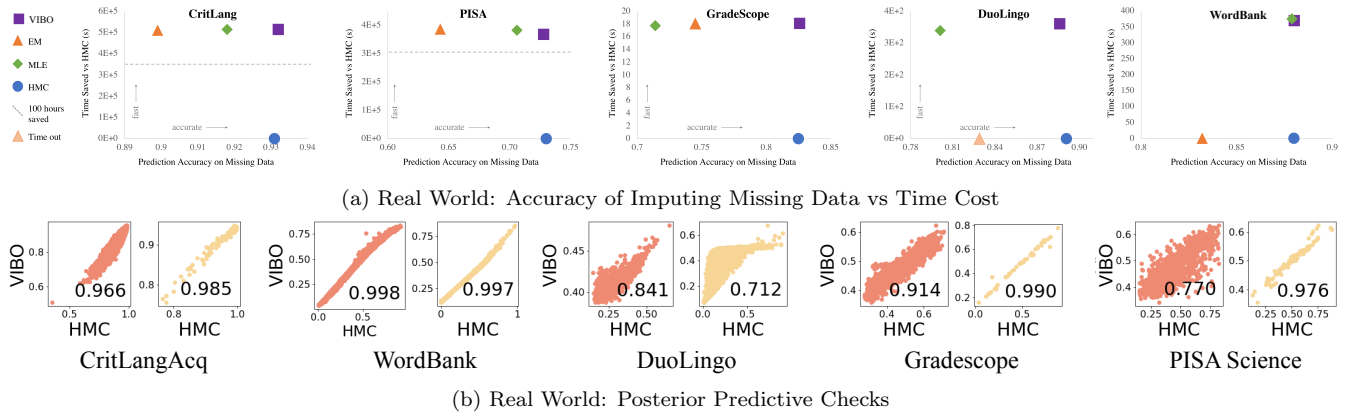(b) Real World: Posterior Predictive Checks

Figure 3: (a) Accuracy of missing data imputation for real world datasets plotted against time saved in seconds compared to using HMC. (b) Samples statistics from the predictive posterior defined using HMC and VIBO. A correlation of 1.0 would be perfect alignment between two inference techniques. Subfigures in red show the average number of items answered correctly for each person. Subfigures in yellow show the average number of people who answered each item correctly.

Fig. 2 shows run-time and performance results for VIBO, MLE, HMC, EM, and two ablations of VIBO (discussed in Sec. 5.4). First, comparing parameter recovery performance (Fig. 2 middle), we see that HMC, MLE and VIBO all recover parameters well. The only notable differences are: VIBO with very few people, and HMC and (to a lesser extent) VIBO in high dimensions. The former is because the amortized posterior approximation requires a sufficiently large dataset (around 500 people) to constrain its parameters. The latter is a simple effect of the scaling of variance for sample-based estimates as dimensionality increases (we fixed the number of samples used, to ease speed comparisons).

Turning to the ability to predict missing data (Fig. 2 bottom) we see that VIBO performs equally well to HMC, except in the case of very few people (again, discussed below). (Note that the latent dimensionality does not adversely affect VIBO or HMC for missing data prediction, because the variance is marginalized away.) MLE also performs well as we scale number of items and latent ability dimensions, but is less able to benefit from more people. EM on the other hand provides much worse missing data prediction in all cases.

Finally if we examine the speed of inference (Fig. 2 top), VIBO is only slightly slower than MLE, both of which are orders of magnitude faster than HMC. For instance, with 1.56 million people, HMC takes 217 hours whereas VIBO takes 800 seconds. Similarly with 6250 items, HMC takes 4.3 hours whereas VIBO takes 385 seconds. EM is the fastest for low to medium sized datasets, though its lower accuracy makes this a dubious victory. Furthermore, EM does not scale as well as VIBO to large datasets.

## 5.3 Real World Data Results

We next apply VIBO to real world datasets in cognitive science and education. Fig. 3(a) plots the accuracy of imputing missing data against the time saved vs HMC (the most expensive inference algorithm) for five large-scale datasets. Points in the upper right corner are more desirable as they are more accurate and faster. The dotted line represents 100 hours saved compared to HMC.

From Fig. 3(a), we find many of the same patterns as we observed in the synthetic experiments. Running HMC on CritLangAcq or PISA takes roughly 120 hours whereas VIBO takes 50 minutes for CritLangAcq and 5 hours for PISA, the latter being more expensive because of computation required for missing data. In comparison, EM is at times faster than VIBO (e.g. Gradescope, PISA) and at times slower. With respect to accuracy, VIBO and HMC are again identical, outperforming EM by up to 8% in missing data imputation. Interestingly, we find the "overfitting" of MLE to be more pronounced here. If we focus on DuoLingo and Gradescope, the two datasets with pre-existing large portions of missing values, MLE is surpassed by EM, with VIBO achieving accuracies 10% higher.

Another way of exploring a model's ability to explain data, for fully Bayesian models, is posterior predictive checks. Fig. 3(b) shows posterior predictive checks comparing VIBO and HMC. We find that the two algorithms strongly agree about the average number of correct people and items in all datasets. The only systematic deviations occur with DuoLingo: it is possible that this is a case where a more expressive posterior approximation would be useful in VIBO, since the number of items is greater than the number of people.

## 5.4 Ablation Studies

We compared VIBO to simpler variants that either do not amortize the posterior or do so with independent distributions of ability and item parameters. These correspond to different variational families, $\mathcal{Q}$ to choose $q$ from:

- VIBO (Independent): We consider the decomposition $q(\boldsymbol{a}_i, \boldsymbol{d}_{1:M} | \boldsymbol{r}_{i,1:M}) = q(\boldsymbol{a}_i | \boldsymbol{r}_{i,1:M}) q(\boldsymbol{d}_{1:M})$ which treats ability and item characteristics as independent.

- VIBO (Unamortized): We consider $q(\boldsymbol{a}_i, \boldsymbol{d}_{1:M} | \boldsymbol{r}_{i,1:M}) = q_{\psi(\boldsymbol{r}_{i,1:M})}(\boldsymbol{a}_i) q(\boldsymbol{d}_{1:M})$, which learns separate posteriors for each $\boldsymbol{a}_i$, without parameter sharing. Recall the subscripts $\psi(\boldsymbol{r}_{i,1:M})$ indicate a separate variational posterior for each unique set of responses.

If we compare unamortized to amortized VIBO in Fig. 2 (top), we see an important efficiency difference. The number

of parameters for the unamortized version scales with the number of people; the speed shows a corresponding impact, with the amortized version becoming an order of magnitude faster than the unamortized one. In general, amortized inference is much cheaper, especially in circumstances in which the number of possible response vectors $r_{1:M}$ is very large (e.g. $2^{95}$ for CritLangAcq). Comparing amortized VIBO to the un-amortized equivalent, Table 2 compares the wall clock time (sec.) for the 5 real world datasets. While VIBO is comparable to MLE and EM (Fig. 3a), unamortized VIBO is 2 to 15 times more expensive.

Exploring accuracy in Fig. 2 (bottom), we see that the unamortized variant is significantly less accurate at predicting missing data. This can be attributed to overfitting to observed responses. With 100 items, there are $2^{100}$ possible responses from every person, meaning that even large datasets only cover a small portion of the full set. With amortization, overfitting is more difficult as the deterministic mapping $f_\phi$ is not hardcoded to a single response vector. Without amortization, since we learn a variational posterior for every observed response vector, we may not generalize to new response vectors. Unamortized VIBO is thus much more sensitive to missing data as it does not get to observed the entire response. We can see evidence of this as unamortized VIBO is superior to amortized VIBO at parameter recovery, Fig. 2 (middle), where no data is hidden from the model; compare this to missing data imputation, where unamortized VIBO appears inferior: because ability estimates do not share parameters, those with missing data are less constrained yielding poorer predictive performance.

Finally, when there are very few people (100) unamortized VIBO and HMC are better at recovering parameters (Fig. 2 middle) than amortized VIBO. This can be explained by amortization: to train an effective regressor $f_\phi$ requires a minimum amount of data. With too few responses, the amortization gap will be very large, leading to poor inference. Under scarce data we would thus recommend using HMC, which is fast enough and most accurate.

Table 2: Time Costs with and without Amortization

| Dataset | Amortized (Sec.) | Un-Amortized (Sec.) |
|---|---|---|
| CritLangAcq | 2.8k | 43.2k |
| WordBank | 176.4 | 657.1 |
| DuoLingo | 429.9 | 717.9 |
| Gradescope | 114.5 | 511.1 |
| PISA | 25.2k | 125.8k |

The above suggests that amortization is important when dealing with moderate to large datasets. Turning to the structure of the amortized posteriors, we note that the factorization we chose in Thm. 3.1 is only one of many. Specifically, we could make the simpler assumption of independence between ability and item characteristics given responses in our variational posteriors: VIBO (Independent). Such a factorization would be simpler and faster due to less gradient computation. However, in our synthetic experiments (in which we know the true ability and item features), we found the independence assumption to produce very poor results: recovered ability and item characteristics had less than 0.1 correlation with the true parameters. Meanwhile

the factorization we posed in Thm. 3.1 consistently produced above 0.9 correlation. Thus, the insight to decompose $q(a_i, d_{1:M}|r_{i,1:M}) = q(a_i|d_{1:M}, r_{i,1:M})q(d_{1:M}|r_{i,1:M})$ instead of assuming independence is a critical one. (This point is also supported theoretically by research on faithful inversions of graphical models [43].)

# 6. DEEP ITEM RESPONSE THEORY

We have found VIBO to be fast and accurate for inference in 2PL IRT, matching HMC in accuracy and EM in speed. This classic IRT model is a surprisingly good model for item responses despite its simplicity. Yet it makes strong assumptions about the interaction of factors, which may not capture the nuances of human cognition. With the advent of much larger data sets we have the opportunity to explore corrections to classic IRT models, by introducing more flexible non-linearities. As described above, a virtue of VI is the possibility of learning aspects of the generative model by optimizing the inference objective. We next explore several ways to incorporate learnable non-linearities in IRT, using the modern machinery of deep learning.

## 6.1 Nonlinear Generalizations of IRT

We have assumed thus far that $p(r_{i,1:M}|a_i, d_{1:M})$ is a fixed IRT model defining the probability of correct response to each item. We now consider three different alternatives with varying levels of expressivity that help define a class of more powerful nonlinear IRT.

**Learning a Linking Function** We replace the logistic function in standard IRT with a nonlinear linking function. As such, it preserves the linear relationships between items and people. We call this VIBO (Link). For person $i$ and item $j$, the 2PL-Link generative model is:

$$p(r_{ij}|a_i, d_j) = f_\theta(-a_i^T k_j - d_j) \tag{16}$$

where $f_\theta$ is a one-dimensional nonlinear function followed by a sigmoid to constrain the output to be within $[0, 1]$. In practice, we parameterize $f_\theta$ as a multilayer perceptron (MLP) with three layers of 64 hidden nodes with ELU nonlinearities.

**Learning a Neural Network** Here, we no longer preserve the linear relationships between items and people and instead feed the ability and item characteristics directly into a neural network, which will combine the inputs nonlinearly. We call this version VIBO (Deep). For person $i$ and item $j$, the Deep generative model is:

$$p(r_{ij}|a_i, d_j) = f_\theta(a_i, d_j) \tag{17}$$

where again $f_\theta$ includes a Sigmoid function at the end to preserve the correct output signatures. This is an even more expressive model than VIBO (Link). In practice, we parameterize $f_\theta$ as three MLPs, each with 3 layers of 64 nodes and ELU nonlinearities. The first MLP maps ability to a real vector; the second maps item characteristics to a real vector. These two hidden vectors are concatenated and given to the final MLP, which predicts response.

**Learning a Residual Correction** Although clearly a powerful model, we might fear that VIBO (Deep) becomes too uninterpretable. So, for the third and final nonlinear model, we use the standard IRT but add a nonlinear residual

component that can correct for any inaccuracies. We call this version VIBO (Residual). For person $i$ and item $j$, the 2PL-Residual generative model is:

$$p(r_{ij}|\boldsymbol{a}_i, \boldsymbol{k}_j, d_j) = \frac{1}{1 + e^{-\boldsymbol{a}_i^T \boldsymbol{k}_j - d_j + f_\theta(\boldsymbol{a}_i, \boldsymbol{k}_j, d_j)}} \quad (18)$$

During optimization, we initialize the weights of the residual network to 0, thus ensuring its initial output is 0. This encourages the model to stay close to IRT, using the residual only when necessary. We use the same architectures for the residual component as in VIBO (Deep).

## 6.2 Nonlinear IRT Evaluation

A generative model explains the data better when it assigns observations higher probability. We thus evaluate generative models by estimating the log marginal likelihood $\log p(\boldsymbol{r}_{1:N,1:M})$ of the training dataset. A higher number (closer to 0) is better. For a single person, the log marginal likelihood of his or her $M$ responses can be computed as:

$$\log p(\boldsymbol{r}_{i,1:M}) \approx \log \mathbb{E}_{q_\phi(\boldsymbol{a}_i, \boldsymbol{d}_{1:M}|\boldsymbol{r}_{i,1:M})} \left[ \frac{p_\theta(\boldsymbol{r}_{i,1:M}, \boldsymbol{a}_i, \boldsymbol{d}_{1:M})}{q_\phi(\boldsymbol{a}_i, \boldsymbol{d}_{1:M}|\boldsymbol{r}_{i,1:M})} \right]$$
$$(19)$$

We use 1000 samples to estimate Eq. 19. We also measure accuracy on missing data imputation as we did in Sec. 5. A more powerful generative model, that is more descriptive of the data, should be better at filling in missing values.

## 6.3 Nonlinear IRT Results

The top half of Table 3 compares the log likelihoods of observed data whereas the bottom half of Table 3 compares the accuracy of imputing missing data. We include VIBO inference with classical IRT-1PL and IRT-2PL generative models as baselines. We find a consistent trend: the more powerful generative models achieve a higher log likelihood (closer to 0) and a higher accuracy. In particular, we find very large increases in log likelihood moving from IRT to Link, spanning 100 to 500 log points depending on the dataset. Further, from Link to Deep and Residual, we find another increase of 100 to 200 log points. In some cases, we find Residual to outperform Deep, though the two are equally parameterized, suggesting that initialization with IRT can find better local optima. These gains in log likelihood translate to a consistent 1 to 2% increase in held-out accuracy for Link/Deep/Residual over IRT. This suggests that the datasets are large enough to use the added model flexibility appropriately, rather than overfitting to the data.

We also compare our deep generative IRT models with the purely deep learning approach called Deep-IRT [47] (see Sec. 8), that does not model posterior uncertainty. Unlike traditional IRT models, Deep-IRT was built for knowledge tracing and assumed sequential responses. To make our datasets amenable to Deep-IRT, we assume an ordering of responses from $j = 1$ to $j = M$. As shown in Table 3, our models outperform Deep-IRT in all 5 datasets by as much as 30% in missing data imputation (e.g. WordBank).

## 6.4 Interpreting the Linking Function

With nonlinear models, we face an unfortunate tradeoff between interpretability and expressivity. In domains like education, practitioners greatly value the interpretability of IRT where predictions can be directly attributed to ability or item features. With VIBO (Deep), our most expressive model, predictions use a neural network, making it hard to understand the interactions between people and items.
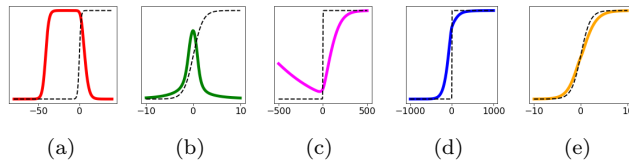


Figure 4: Learned link functions for (a) CritLangAcq, (b) WordBank, (c) DuoLingo, (d) Gradescope, and (e) PISA. The dotted black line shows the default logistic function.

Fortunately, with VIBO (Link), we can maintain a degree of interpretability along with power. The "Link" generative model is identical to IRT, only differing in the linking function (i.e. item response function). Each subfigure in Fig. 4 shows the learned response function for one of the real world datasets; the dotted black line represents the best standard linking function, a sigmoid. We find three classes of linking functions: (1) for Gradescope and PISA, the learned function stays near a Sigmoid. (2) For WordBank and CritLangAcq, the response function closely resembles an unfolding model [27, 2], which encodes a more nuanced interaction between ability and item characteristics: higher scores are related to higher ability only if the ability and item characteristics are "nearby" in latent space. (3) For DuoLingo, we find a piecewise function that resembles a sigmoid for positive values and a negative linear function for negative values. In cases (2) and (3) we find much greater differences in log likelihood between VIBO (IRT) and VIBO (Link). See Table 3. For DuoLingo, VIBO (Link) matches the log density of more expressive models, suggesting that most of the benefit of nonlinearity is exactly in this unusual linking function.

## 7. POLYTOMOUS RESPONSES

Thus far, we have been working only with response data collapsed into binary correct/incorrect responses. However, many questionnaires and examinations are not binary: responses can be multiple choice (e.g. Likert scale) or even real valued (e.g. 92% on a course test). Having posed IRT as a generative model, we have prescribed a Bernoulli distribution over the $i$-th person's response to the $j$-th item. Yet nothing prevents us from choosing a different distribution, such as Categorical for multiple choice or Normal for real-values. The DuoLingo dataset contains partial credit, computed as a fraction of times an individual gets a word correct. A more granular treatment of these polytomous values should yield a more faithful model that can better capture the differences between people. We thus modeled the DuoLingo data using for $p(\boldsymbol{r}_{i,1:M}|\boldsymbol{a}_i, \boldsymbol{d}_{1:M})$ a (truncated) Normal distribution over responses with fixed variance. Table 4 show the log densities: we again observe large improvements from nonlinear models.

Item Response Theory can in this way be extended to work of all kinds (imagine students writing text, drawing pictures, or even coding), encouraging educators to assign open-ended work without having to give up proper tools of assessment.

## 8. RELATED WORK

Table 3: Log Likelihoods and Missing Data Imputation for Deep Generative IRT Models

| DATASET | DEEP IRT | VIBO (IRT-1PL) | VIBO (IRT-2PL) | VIBO (LINK-2PL) | VIBO (DEEP-2PL) | VIBO (RES.-2PL) |
|---|---|---|---|---|---|---|
| CRITLANGACQ | - | $-11249.8 \pm 7.6$ | $-10224.0 \pm 7.1$ | $-9590.3 \pm 2.1$ | $-9311.2 \pm 5.1$ | $\mathbf{-9254.1} \pm 4.8$ |
| WORDBANK | - | $-17047.2 \pm 4.3$ | $-5882.5 \pm 0.8$ | $-5268.0 \pm 7.0$ | $\mathbf{-4658.4} \pm 3.9$ | $-4681.4 \pm 2.2$ |
| DUOLINGO | - | $-2833.3 \pm 0.7$ | $-2488.3 \pm 1.4$ | $-1833.9 \pm 0.3$ | $-1834.2 \pm 1.3$ | $\mathbf{-1745.4} \pm 4.7$ |
| GRADESCOPE | - | $-1090.7 \pm 2.9$ | $-876.7 \pm 3.5$ | $-750.8 \pm 0.1$ | $\mathbf{-705.1} \pm 0.5$ | $-715.3 \pm 2.7$ |
| PISA | - | $-13104.2 \pm 5.1$ | $-6169.5 \pm 4.8$ | $-6120.1 \pm 1.3$ | $-6030.2 \pm 3.3$ | $\mathbf{-5807.3} \pm 4.2$ |
| CRITLANGACQ | 0.934 | 0.927 | 0.932 | 0.945 | **0.948** | 0.947 |
| WORDBANK | 0.681 | 0.876 | 0.880 | 0.888 | 0.889 | **0.889** |
| DUOLINGO | 0.884 | 0.880 | 0.886 | 0.891 | **0.897** | 0.894 |
| GRADESCOPE | 0.813 | 0.820 | 0.826 | 0.840 | 0.847 | **0.848** |
| PISA | 0.524 | 0.723 | 0.728 | 0.718 | **0.744** | 0.739 |

Table 4: DuoLingo with Polytomous Responses

| INF. ALG. | TRAIN | TEST |
|---|---|---|
| VIBO (IRT) | $-22038.07$ | $-21582.03$ |
| VIBO (LINK) | $-17293.35$ | $-16588.06$ |
| VIBO (DEEP) | $\mathbf{-15349.84}$ | $\mathbf{-14972.66}$ |
| VIBO (RES.) | $-15350.66$ | $-14996.27$ |

We described above a variety of methods for parameter estimation in IRT such as MLE, EM, and MCMC. The benefits and drawbacks of these methods are well-documented [26], so we need not discuss them here. Instead, we focus specifically on methods that utilize deep neural networks or variational inference to estimate IRT parameters.

While variational inference has been suggested as a promising alternative to other inference approaches for IRT [26], there has been surprisingly little work in this area. In an exploration of Bayesian prior choice for IRT estimation, Natesan et al. [31] posed a variational approximation to the posterior:

$$p(\boldsymbol{a}_i, \boldsymbol{d}_j | r_{i,j}) \approx q_\phi(\boldsymbol{a}_i, \boldsymbol{d}_j) = q_\phi(\boldsymbol{a}_i)q_\phi(\boldsymbol{d}_j) \qquad (20)$$

This is an unamortized and independent posterior family, unlike VIBO. As we noted in Sec. 5.4, both amortization and dependence of ability on items were crucial for our results.

We are aware of two approaches that incorporate deep neural networks into Item Response Theory: Deep-IRT [46] and DIRT [8]. Deep-IRT is a modification of the Dynamic Key-Value Memory Network (DKVMN) [47] that treats data as longitudinal, processing items one-at-a-time using a recurrent architecture. Deep-IRT produces point estimates of ability and item difficulty at each time step, which are then passed into a 1PL IRT function to produce the probability of answering the item correctly. The main difference between DIRT and Deep-IRT is the choice of neural network: instead of the DKVMN, DIRT uses an LSTM with attention [41]. In our experiments, we compare our approach to Deep-IRT and find that we outperform it by up to 30% on the accuracy of missing response imputation. On the other hand, our models do not capture the longitudinal aspect of response data. Combining the two approaches would be natural.

Lastly, Curi et al. [9] used a VAE to estimate IRT parameters in a 28-question synthetic dataset. However, this approach modeled ability as the only unknown variable, ignoring items. Our analogue to the VAE builds on the IRT graphical model, incorporating both ability and item characteristics in a principled manner. This could explain why Curi et. al. report the VAE requiring substantially more data to recover the true parameters when compared to MCMC whereas we find comparable data-efficiency between VIBO and MCMC.

## 9. BROADER IMPACT
We briefly emphasize the broader impact of efficient IRT in the context of education. Firstly, one of the many difficulties of accurately estimating student ability is cost: attempting to use MCMC on the order magnitude required by large entities like MOOCs, local and national governments, and international organizations is impossible. However with VIBO, doing so is already possible, as shown by the PISA results. Second, efficient IRT is an important and necessary step to encourage the development of more complex models of student cognition and response. Namely, it will at least enable faster research and iterative testing on real world data.

## 10. CONCLUSION
Item Response Theory is a paradigm for reasoning about the scoring of tests, surveys, and similar measurment instruments. Notably, the theory plays an important role in education, medicine, and psychology. Inferring ability and item characteristics poses a technical challenge: balancing efficiency against accuracy. In this paper we have found that variational inference provides a potential solution, running orders of magnitude faster than MCMC algorithms while matching their state-of-the-art accuracy.

Many directions for future work suggest themselves. First, further gains in speed and accuracy could be found by exploring more or less complex families of posterior approximation. Second, more work is needed to understand deep generative IRT models and determine the most appropriate tradeoff between expressivity and interpretability. For instance, we found significant improvements from a learned linking function, yet in some applications monotonicity may be judged important to maintain – greater ability, for instance, should correspond to greater chance of success. Finally, VIBO should enable more coherent, fully Bayesian, exploration of very large and important datasets, such as PISA [13].

Recent advances within AI combined with new massive datasets have enabled advances in many domains. We have given an example of this fruitful interaction for understanding humans based on their answers to questions.

# 11. ACKNOWLEDGMENTS

# 12. REFERENCES

[1] T. A. Ackerman. Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education*, 7(4):255–278, 1994.

[2] D. Andrich and G. Luo. A hyperbolic cosine latent trait model for unfolding dichotomous single-stimulus responses. *Applied Psychological Measurement*, 17(3):253–276, 1993.

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

[4] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

[5] R. D. Bock and M. Aitkin. Marginal maximum likelihood estimation of item parameters: Application of an em algorithm. *Psychometrika*, 46(4):443–459, 1981.

[6] R. D. Bock, R. Gibbons, and E. Muraki. Full-information item factor analysis. *Applied Psychological Measurement*, 12(3):261–280, 1988.

[7] M. Braginsky, D. Yurovsky, V. A. Marchman, and M. C. Frank. Developmental changes in the relationship between grammar and the lexicon. In *CogSci*, pages 256–261, 2015.

[8] S. Cheng, Q. Liu, E. Chen, Z. Huang, Z. Huang, Y. Chen, H. Ma, and G. Hu. Dirt: Deep learning enhanced item response theory for cognitive diagnosis. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 2397–2400, 2019.

[9] M. Curi, G. A. Converse, J. Hajewski, and S. Oliveira. Interpretable variational autoencoders for cognitive models. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019.

[10] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[11] F. Y. Edgeworth. The statistics of examinations. *Journal of the Royal Statistical Society*, 51(3):599–635, 1888.

[12] S. E. Embretson and S. P. Reise. *Item response theory*. Psychology Press, 2013.

[13] O. for Economic Co-operation and D. (OECD). Pisa 2015 database. 2016.

[14] M. C. Frank, M. Braginsky, D. Yurovsky, and V. A. Marchman. Wordbank: An open repository for developmental vocabulary data. *Journal of child language*, 44(3):677–694, 2017.

[15] A. E. Gelfand and A. F. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398–409, 1990.

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

[17] S. J. Haberman. Maximum likelihood estimates in exponential response models. *The Annals of Statistics*, pages 815–841, 1977.

[18] W. Harlen. The assessment of scientific literacy in the oecd/pisa project. 2001.

[19] J. K. Hartshorne, J. B. Tenenbaum, and S. Pinker. A critical period for second language acquisition: Evidence from 2/3 million english speakers. *Cognition*, 177:263–277, 2018.

[20] M. R. Harwell, F. B. Baker, and M. Zwarts. Item parameter estimation via marginal maximum likelihood and an em algorithm: A didactic. *Journal of Educational Statistics*, 13(3):243–271, 1988.

[21] W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. 1970.

[22] G. E. Hinton. Products of experts. 1999.

[23] M. D. Hoffman and A. Gelman. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15(1):1593–1623, 2014.

[24] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

[25] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[26] W. J. v. d. Linden. *Handbook of Item Response Theory: Volume 2: Statistical Tools*. CRC Press, 2017.

[27] C.-W. Liu and R. P. Chalmers. Fitting item response unfolding models to likert-scale data using mirt in r. *PloS one*, 13(5), 2018.

[28] L. Magdalena, E. Haman, S. A. Lotem, B. Etenkowski, F. Southwood, D. Andjelkovic, E. Bloom, T. Boerma, S. Chiat, P. E. de Abreu, et al. Ratings of age of acquisition of 299 words across 25 languages: Is there a cross-linguistic order og words? *Behavior Research Methods (print Edition)*, 48(3):1154–1177, 2016.

[29] R. P. McDonald. A basis for multidimensional item response theory. *Applied Psychological Measurement*, 24(2):99–114, 2000.

[30] A. Mnih and K. Gregor. Neural variational inference and learning in belief networks. *arXiv preprint arXiv:1402.0030*, 2014.

[31] P. Natesan, R. Nandakumar, T. Minka, and J. D. Rubright. Bayesian prior choice in irt estimation using mcmc and variational bayes. *Frontiers in psychology*, 7:1422, 2016.

[32] R. M. Neal. An improved acceptance procedure for the hybrid monte carlo algorithm. *Journal of Computational Physics*, 111(1):194–203, 1994.

[33] R. M. Neal et al. Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2, 2011.

[34] G. Rasch. Studies in mathematical psychology: I. probabilistic models for some intelligence and attainment tests. 1960.

[35] D. Ravitch. *National standards in American education:*

*A citizen's guide.* ERIC, 1995.

[36] M. D. Reckase. Multidimensional item response theory models. In *Multidimensional item response theory*, pages 79–112. Springer, 2009.

[37] D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

[38] B. Settles, C. Brust, E. Gustafson, M. Hagiwara, and N. Madnani. Second language acquisition modeling. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, pages 56–65, 2018.

[39] A. Singh, S. Karayev, K. Gutowski, and P. Abbeel. Gradescope: a fast, flexible, and fair system for scalable assessment of handwritten work. In *Proceedings of the fourth (2017) acm conference on learning@ scale*, pages 81–88. ACM, 2017.

[40] S. Sinharay, M. S. Johnson, and H. S. Stern. Posterior predictive assessment of item response theory models. *Applied Psychological Measurement*, 30(4):298–321, 2006.

[41] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[42] M. J. Wainwright, M. I. Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

[43] S. Webb, A. Golinski, R. Zinkov, S. Narayanaswamy, T. Rainforth, Y. W. Teh, and F. Wood. Faithful inversion of generative models for effective amortized inference. In *Advances in Neural Information Processing Systems*, pages 3070–3080, 2018.

[44] R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

[45] M. Wu and N. Goodman. Multimodal generative models for scalable weakly-supervised learning. In *Advances in Neural Information Processing Systems*, pages 5575–5585, 2018.

[46] C.-K. Yeung. Deep-irt: Make deep learning based knowledge tracing explainable using item response theory. *arXiv preprint arXiv:1904.11738*, 2019.

[47] J. Zhang, X. Shi, I. King, and D.-Y. Yeung. Dynamic key-value memory networks for knowledge tracing. In *Proceedings of the 26th international conference on World Wide Web*, pages 765–774, 2017.