Psychometrika (2026), **0**, 1–19
doi:10.1017/psy.2026.10105

PSYCHOMETRIC SOCIETY

APPLICATION AND CASE STUDIES - ORIGINAL

Estimating Discrete Latent Variable Models Using Amortized Variational Inference

Karel Veldkamp[1], Raoul Grasman[1] and Dylan Molenaar[2]

[1]Psychology, Universiteit van Amsterdam, Netherlands; [2]Faculty of Social and Behavioural Sciences, University of Amsterdam, Netherlands

Corresponding author: Karel Veldkamp; Email: k.a.veldkamp@uva.nl

(Received 26 July 2025; revised 23 February 2026; accepted 16 March 2026)

Abstract

Recent research shows that amortized variational inference (AVI) can be used to efficiently estimate high-dimensional latent variable models on large datasets. However, its use has remained limited to item response theory (IRT), and generalizing the approach to discrete latent variable models is not straightforward. We propose two ways to deal with this problem. In an initial simulation, we verify that these approaches can be used to estimate simple discrete latent variable models, such as latent class analysis and the generalized deterministic inputs, noisy and gate model. In these cases, AVI provides accurate parameter estimates, although the computational advantage over marginal maximum likelihood (MML) and standard variational inference (VI) is limited. We then apply the same approach to estimate mixture IRT models. In this case, AVI is computationally faster than MML estimation and standard VI. To demonstrate the practical applicability of our AVI approach, we use it to fit a seven-dimensional mixture IRT model to a narcissism inventory. Whereas quadrature-based methods cannot feasibly estimate models of this dimensionality, the efficient AVI approach even allows for computation of bootstrapped standard errors. We provide our code, along with an easy-to-use tool for fitting these models to new datasets.

Keywords: amortized variational inference; discrete latent variables; high-dimensional models; latent class analysis; mixture IRT

1. Introduction

Latent variable models are widely used in psychometrics to explain the covariance between a set of observed variables by a set of unobserved, or latent variables. The most common way to fit latent variable models is by marginal maximum likelihood (MML) using the expectation–maximization (EM) algorithm, which is attractive due to its asymptotic properties. However, when the number of latent dimensions in the model increases, MML can become computationally infeasible due to the high-dimensional nature of the integral in the marginal likelihood. This computational burden is further amplified by larger sample sizes, which increase the amount of data over which these integrals have to be evaluated. Yet, large, high-dimensional datasets are becoming more and more common due to the ease of online data collection, the rise of educational and psychological monitoring systems, and the increased availability of process data. Various solutions to this computational problem exist in the literature, such as adaptive quadrature (Bock & Aitkin, 1981), Markov chain Monte Carlo (MCMC) (Edwards, 2010), stochastic approximation (Cai, 2010; von Davier & Sinharay, 2010), and joint

maximum likelihood (Chen et al., 2019). However, fitting high-dimensional models to large datasets is generally still computationally challenging.

An alternative to MML that has recently been developed to estimate latent variable models is standard variational inference (VI). These methods approximate the posterior distribution of the latent variables by a highly factorial distribution. By sampling from this approximate distribution, a lower bound on the marginal likelihood can be maximized to obtain parameter estimates (Blei et al., 2017). This method was originally developed in a Bayesian context and has been used to estimate Bayesian item response theory (IRT) (Wu et al., 2021) and cognitive diagnostic (Yamaguchi & Okada, 2020) models successfully. However, the same approach has also been used to estimate models with fixed item parameters (Cho et al., 2021). Furthermore, by taking multiple samples from the approximate posterior using importance-weighted sampling (Kahn & Marshall, 1953), the lower bound can be made arbitrarily close to the marginal likelihood depending on the number of samples, approaching the MML solution (Ma et al., 2023).

Although VI avoids the evaluation of the integral in the marginal likelihood, the posterior parameters for each individual must be estimated. The number of variational parameters scales with the sample size N, which becomes computationally intensive for large datasets and high-dimensional models, especially if there is no closed-form update possible for the approximate posteriors. This drawback is avoided in amortized VI (AVI), where instead of directly optimizing the posterior parameters, a so-called inference model is estimated. This inference model relates the observed data to the posterior parameters using some non-linear parametric function. That is, the posterior parameters are *amortized*. Note that this is different from Bayesflow style amortization (Radev et al., 2020), which relies on large-scale simulation to train a network that maps data to posterior distributions. In AVI, amortization refers only to predicting parameters of the approximate posteriors from the observed data, not relying on any simulation. During estimation, the parameters of the approximate posterior are computed using the current estimates from the inference model, a sample is taken from the approximate posterior, and the parameters of the measurement and inference models are updated jointly using gradient descent. The use of amortization improves the estimation efficiency on large datasets and high-dimensional models as the number of parameters in the inference model depend on the number of items but not on the number of individuals. Curi et al. (2019) were one of the first to use such an approach to estimate IRT parameters, and Urban and Bauer (2021) extended this approach to include multiple importance-weighted samples from the approximate posterior to approximate the marginal likelihood solution even closer. In both studies, it is demonstrated that for large, high-dimensional data, this method of estimation is much faster than state-of-the-art MML implementations, while still attaining approximately equally accurate parameter estimates. Since then, the AVI approach has been applied to a broad range of IRT models. For example, the approach has been generalized to the three- and four-parameter logistic IRT models (Liu et al., 2022), models that allow for correlated latent variables (Converse et al., 2021), fixed effect IRT models (Molenaar et al., 2025), models that jointly model responses and response times (Debelak & Urban, 2021), and models that amortize over both person and item parameters (Zhang & Chen, 2024). Additionally, the model has been generalized to situations where the factor loading structure is misspecified (Hasan et al., 2022) and to models for data with a significant proportion of missingness (Veldkamp et al., 2024).

When looking at the taxonomy of latent variable models by Bartholomew et al. (2011) in Table 1, it is interesting to note that currently, the applications of AVI mostly—if not exclusively—focus on IRT models. Although it is straightforward to generalize the AVI approach to factor analysis, the benefits are less prominent as there is a closed-form solution for the integrals in the marginal likelihood of the (normal theory) linear factor model. A more interesting application for AVI concerns models with discrete latent variables, such as latent class and latent profile models. Such discrete latent variable models have many important applications in the fields of educational psychology (de la Torre & Minchen, 2014), clinical psychology (Sullivan et al., 1998), medicine (Formann & Kohlmann, 1996), and sociology (Yamaguchi, 2000). Although there is no closed-form solution for the integrals in the marginal likelihood of latent variable models including latent classes, the computational advantage of

https://doi.org/10.1017/psy.2026.10105 Published online by Cambridge University Press

Table 1. Four types of latent variable models according to the taxonomy of Bartholomew et al. (2011)

	Discrete latent variables	Continuous latent variables
Discrete observed variables	Latent class models (LCMs)	Item response theory (IRT)
Continuous observed variables	Latent profile analysis	Factor analysis

AVI for simple latent class or latent profile models is expected to be limited. That is, evaluating the integrals in the marginal likelihood of a discrete latent variable model is computationally less demanding as the parameter space to be intergrated over is much smaller than that of the continuous latent variables in IRT models.

Although AVI has not been used for discrete latent variables, standard VI has successfully been used to estimate discrete latent variable models. For example, Grimmer (2011) used VI to estimate latent class models and Oka and Okada (2023) and Yamaguchi and Okada (2020) used VI to estimate cognitive diagnostic models. These papers are both fully Bayesian including prior on both the latent variables and the item parameters. However, priors on the item parameters are by no means a requirement for a discrete VI approach, and we estimate fixed item parameters in this work. Importantly, previous work on VI and discrete latent variables use conjugate priors, yielding closed-form analytical updates for the variational posteriors, greatly improving estimation efficiency. However, as the conditional distributions involved need to be part of the exponential family, the generalizability of these approaches is relatively limited.

Hybrid models that include both discrete and continuous latent variables, such as mixture IRT models (Sen & Cohen, 2019) and factor mixture models (Kim et al., 2023), can therefore potentially benefit substantially from AVI, particularly when the observed variables are categorical. In such models, both the discrete and continuous latent variables need to be integrated out of the likelihood, so estimation becomes computationally challenging even with a moderate number of latent variables. Extending the AVI framework to these hybrid models, while avoiding restrictive conjugacy assumptions, is thus a promising direction. However, as discussed below, incorporating discrete latent variables into AVI is not straightforward.

Estimating the parameters of the inference model requires reparameterizing the sampling process into a differentiable function that depends on the parameters from the posterior distribution and a random variable with fixed parameters (Kingma & Welling, 2013). While this is straightforward for continuously distributed latent variables, it is more challenging for discrete latent variables (see Section 3 for a detailed discussion). Several techniques have been proposed in the machine learning literature that allow sampling from discrete distributions. However, while machine learning applications focus on predictive accuracy, latent variable modeling is primarily concerned with accurate parameter estimates, since the parameters themselves carry substantive meaning and are used for (high-stakes) practical decisions. In the current article, we propose two methods to incorporate discrete latent variables in the AVI framework. We provide a systematic comparison of these methods in estimating latent class models and cognitive diagnostic models. Then, we demonstrate how the most promising method can be generalized to more complex, hybrid latent variable models such as a multidimensional mixture latent variable model, and demonstrate its computational advantage.

The outline is as follows: First, we formally introduce AVI and discuss the challenges related to discrete latent variables. Next, we propose two solutions. We evaluate these solutions in a preliminary simulation devoted to latent class analysis (LCA) and generalized deterministic inputs, noisy and gate (GDINA) models, verifying that AVI can be used for discrete latent variables. In our main simulation study, we apply the methods for discrete latent variables to estimate a mixture multidimensional IRT model, and demonstrate its computational advantage compared to MML. We then apply this mixture IRT method to a real seven-dimensional large-scale dataset pertaining to Narcissism. We end with a general discussion.

https://doi.org/10.1017/psy.2026.10105 Published online by Cambridge University Press

2. Amortized variational inference

Consider a general latent variable model in which the latent variable vector is denoted by $\boldsymbol{\theta} \in \mathbb{R}^d$ and the corresponding observed response vector is denoted by $\mathbf{x}$. Note that $\boldsymbol{\theta}$ may be continuous or categorical depending on the type of model. The domain will always be clear from the context. The joint distribution of $\boldsymbol{\theta}$ and $\mathbf{x}$ can be written as

$$p_\psi(\boldsymbol{\theta}, \mathbf{x}) = p_\psi(\mathbf{x} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta}), \tag{1}$$

where $p_\psi(\mathbf{x} \mid \boldsymbol{\theta})$ represents the likelihood or measurement model parameterized by (fixed effects) parameters ψ, and $p(\boldsymbol{\theta})$ denotes the population distribution of the (random effects) latent variables. The corresponding posterior distribution of the latent variables given the observed responses is then

$$p_\psi(\boldsymbol{\theta} \mid \mathbf{x}) = \frac{p_\psi(\mathbf{x} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{p_\psi(\mathbf{x})}, \tag{2}$$

where $p_\psi(\mathbf{x}) = \mathbb{E}_{p(\boldsymbol{\theta})}[p_\psi(\mathbf{x} \mid \boldsymbol{\theta})]$ is the marginal likelihood. Each individual i has their own latent variable vector $\boldsymbol{\theta}_i$ and posterior $p_\psi(\boldsymbol{\theta}_i | \mathbf{x}_i)$, but for notational clarity, we drop the index i and write all expressions for a generic observation $(\mathbf{x}, \boldsymbol{\theta})$. In latent variable models for categorical data, this posterior can be computationally expensive to compute or even intractable in the case of continuous latent variables, motivating the use of VI methods.

As already mentioned above, in standard VI, the true posterior distribution is approximated by an approximate posterior distribution which typically (but not necessarily) is a factorized normal distribution. We denote this variational approximation by $q(\boldsymbol{\theta} \mid \mathbf{x})$, meaning the approximate posterior over the latent variable vector $\boldsymbol{\theta} \in \mathbb{R}^d$ given the response vector $\mathbf{x}$. The idea behind VI is to choose the approximate posterior $\hat{q} \in \mathcal{F}$ so that the Kullback–Leibler (KL) divergence to the true posterior, $p_\psi(\boldsymbol{\theta} \mid \mathbf{x})$, is minimized:

$$\hat{q}(\boldsymbol{\theta} \mid \mathbf{x}) = \underset{q(\boldsymbol{\theta}|\mathbf{x}) \in \mathcal{F}}{\operatorname{argmin}} \, \mathrm{KL}[q(\boldsymbol{\theta} \mid \mathbf{x}) \,||\, p_\psi(\boldsymbol{\theta} \mid \mathbf{x})]. \tag{3}$$

Here, $\mathcal{F}$ is a chosen family of allowed approximate posterior distributions. The family $\mathcal{F}$ is usually restricted; for example, to a highly factorized family of distributions. It can be shown that minimizing the KL divergence between the approximate posterior distribution and the trution (Equation (3)) can be accomplished indirectly by maximizing the so-called evidence lower bound (ELBO)

$$\mathrm{ELBO} = \mathbb{E}_{q(\boldsymbol{\theta}|\mathbf{x})}[\log p_\psi(\mathbf{x} \mid \boldsymbol{\theta})] - \mathrm{KL}[q(\boldsymbol{\theta} \mid \mathbf{x}) \,||\, p(\boldsymbol{\theta})], \tag{4}$$

where $p(\boldsymbol{\theta})$ is the prior distribution evaluated at $\boldsymbol{\theta}$ (Blei et al., 2017). The ELBO gives a lower bound to the marginal log-likelihood. It approaches the marginal log-likelihood when the approximate posterior distribution approaches the true posterior distribution. In VI, item parameters ψ are estimated by maximizing this ELBO.

As discussed above, to avoid estimating separate approximate posterior $\hat{q}$ for each individual, in AVI a general, shared inference model q_ζ with parameters ζ is introduced which maps each individual response pattern to its approximate posterior distribution. In other words, instead of optimizing a new variational estimate $\hat{q}$ for every individual, we estimate a single function that models the relation between the approximate posterior parameters and the observed data. Formally, we choose

$$\boldsymbol{\theta} \mid \mathbf{x} \sim q_\zeta(\boldsymbol{\theta} \mid \mathbf{x}), \tag{5}$$

where q_ζ is an inference model with parameters ζ. As such, the approximate posterior is parameterized using parameters ζ common to all individuals. The number of parameters does not depend on the number of individuals. When both the inference model and the measurement model are neural networks, this architecture is also known as a variational autoencoder (VAE) (Kingma & Welling, 2013).

https://doi.org/10.1017/psy.2026.10105 Published online by Cambridge University Press

Although the ELBO is indeed a lower bound to the marginal log-likelihood regardless of the choice of the posterior distribution $q_\zeta(\boldsymbol{\theta} \mid \mathbf{x})$, one often assumes an overly simplified, highly factorized form of $q_\zeta(\boldsymbol{\theta} \mid \mathbf{x})$. Burda et al. (2015) show that this assumption of a simple factorized posterior distribution can lead to a large gap between the ELBO and the marginal likelihood. In order to solve this issue, they propose to use importance-weighted sampling using the approximate posterior distribution as proposal distribution and the joint distribution $p_\psi(\boldsymbol{\theta},\mathbf{x})$ as target distribution. This results in the importance-weighted ELBO (IW-ELBO)

$$\text{IW-ELBO} = \mathbb{E}_{1:S}\left[\log \frac{1}{S}\sum_{s=1}^{S} \frac{p_\psi(\tilde{\boldsymbol{\theta}}^{(s)},\mathbf{x})}{q_\zeta(\tilde{\boldsymbol{\theta}}^{(s)} \mid \mathbf{x})}\right], \tag{6}$$

where S is the number of importance-weighted samples, and $\tilde{\boldsymbol{\theta}}^{(s)} \sim q_\zeta(\boldsymbol{\theta} \mid \mathbf{x})$ are samples drawn from the approximate posterior distribution. By taking the expectation of the joint likelihood using importance-weighted sampling, the IW-ELBO approximates the marginal likelihood more closely. In fact, Burda et al. prove that, as S increases, the IW-ELBO approaches the marginal log-likelihood. Since increasing the number of importance weights can decrease the signal-to-noise ratio of the inference model gradients (Rainforth et al., 2018), it is generally preferable to use doubly reparameterized gradients (DReG), which provide a low-variance unbiased estimate of the IW-ELBO gradients for importance-weighted AVI (Tucker et al., 2018). Urban and Bauer (2021) applied IW-AVI to the estimation of high-dimensional IRT models and showed that taking multiple importance-weighted samples results in parameter estimates that are on par with MML estimates, while these estimates are obtained much more efficiently.

In these AVI approaches, the parameters of the measurement and inference models are optimized jointly using gradient descent. For notational convenience, we rewrite the ELBO in Equation (4) as $\text{ELBO} = \mathbb{E}_{q(\boldsymbol{\theta}|\mathbf{x})}[f_\psi(\boldsymbol{\theta},\mathbf{x})]$, where $f_\psi(\boldsymbol{\theta},\mathbf{x})$ represents the term inside the expectation operator in the ELBO.

Deriving an estimate for the gradients with respect to the parameters of the measurement model ψ is straightforward. Since the expectation is taken over a distribution that is not parameterized by ψ, we can move the gradient inside the expectation:

$$\nabla_\psi \text{ELBO} = \nabla_\psi \mathbb{E}_{q_\zeta(\boldsymbol{\theta}|\mathbf{x})}[f_\psi(\boldsymbol{\theta},\mathbf{x})] = \mathbb{E}_{q_\zeta(\boldsymbol{\theta}|\mathbf{x})}[\nabla_\psi f_\psi(\boldsymbol{\theta},\mathbf{x})], \tag{7}$$

and use L Monte-Carlo samples to approximate the expectation:

$$\approx \frac{1}{L}\sum_{l=1}^{L} \nabla_\psi f_\psi(\tilde{\boldsymbol{\theta}}^{(l)},\mathbf{x}). \tag{8}$$

However, the same is not the case when taking the gradients with respect to ζ since the expectation is taken over $q_\zeta(\boldsymbol{\theta} \mid \mathbf{x})$, which is parameterized by ζ, the gradient cannot simply be moved into the expectation

$$\nabla_\zeta \text{ELBO} = \nabla_\zeta \mathbb{E}_{q_\zeta(\boldsymbol{\theta}|\mathbf{x})}[f_\psi(\boldsymbol{\theta},\mathbf{x})] \neq \mathbb{E}_{q_\zeta(\boldsymbol{\theta}|\mathbf{x})}[\nabla_\zeta f_\psi(\boldsymbol{\theta},\mathbf{x})]. \tag{9}$$

Kingma and Welling (2013) point out that this issue can be avoided by reparameterizing $q_\zeta(\boldsymbol{\theta} \mid \mathbf{x})$ as $p_{\boldsymbol{\varepsilon}}(h_\zeta(\mathbf{x},\boldsymbol{\varepsilon}))$, where h is a deterministic function and $\boldsymbol{\varepsilon}$ is a random sample from a fixed distribution:

$$\nabla_\zeta \text{ELBO} = \nabla_\zeta \mathbb{E}_{p_\varepsilon}[f_\psi(h_\zeta(\mathbf{x},\boldsymbol{\varepsilon}),\mathbf{x})] = \mathbb{E}_{p_\varepsilon}[\nabla_\zeta f_\psi(h_\zeta(\mathbf{x},\boldsymbol{\varepsilon}),\mathbf{x})]. \tag{10}$$

Importantly, there should be an unbiased, low-variance estimate of the gradients of h_ζ in order to effectively estimate ζ using this reparameterization. This has implications for the estimation of models with discrete latent variables, which we discuss below.

https://doi.org/10.1017/psy.2026.10105 Published online by Cambridge University Press

3. AVI for discrete latent variable models

As mentioned above, AVI requires the sample from the posterior latent variable distribution to be parameterized in terms of a differentiable function of its parameters and an independent random variable. Although this is possible for most continuous variables, discrete random variables lack such a differentiable reparameterization (Maddison et al., 2016), making it difficult to use AVI when the latent variables are discrete. Below, we discuss two solutions to this problem.

One solution to this problem is to estimate a continuous vector representation for each class and to estimate a vector of the same dimensions in the inference model. Rather than sampling, the model simply chooses the class that has the vector representation closest to the output of the inference model. More specifically, we introduce a vector of free parameters $\boldsymbol{\eta}_c$ of dimensionality D for each latent class c. We will refer to these vectors as *class embeddings*. The class embeddings are used as input for the measurement model so that the conditional probabilities are conditional on $\boldsymbol{\eta}_c$. The inference model is then defined by the step function

$$q_\zeta(\theta = c|\mathbf{x}) = \begin{cases} 1 & \text{for } c = \underset{c\in\{1,\ldots,C\}}{\arg\min} \; \|e_\zeta(\mathbf{x}) - \boldsymbol{\eta}_c\|_2, \\ 0 & \text{otherwise,} \end{cases} \tag{11}$$

where $e_\zeta(\mathbf{x})$ is the output of the inference model, which is a D-dimensional vector in the same space as $\boldsymbol{\eta}_c$. This is effectively mapping the output of the inference model to the nearest class embedding. Although Equation (11) is non-differentiable, one approach is to use the gradients of the ELBO with respect to the class embeddings and use them as if they are the gradients of the ELBO with respect to the output of the inference model, effectively bypassing the non-differentiable nearest neighbor operation. This is a common approach used to obtain gradients through non-differentiable operations often referred to as straight-through estimators (Bengio et al., 2013). Specifically, we copy the gradients with respect to $\boldsymbol{\eta}_c$ and use them as estimates for the gradients with respect to ζ. The idea to jointly estimate embeddings and an inference model that maps the data onto the nearest embedding is based on research by van den Oord et al. (2017), who uses deep VAEs to determine low-dimensional representations for image, video, and speech generation. We can use a similar embedding architecture to estimate latent class models. We will refer to the resulting model as the *embedding-based latent variable model* (ELVM). Note that while ELVM is well suited for latent class analysis, it is not easy to generalize to more complex models, such as cognitive diagnostic models or mixture models. This is because the class-embedding approach represents each class through an unconstrained continuous vector with no direct interpretability. This works well for latent class models because any unconstrained vector can easily be transformed to probabilities (e.g., a neural network). However for cognitive diagnostic models or mixture models, the probabilities must follow from specific predefined algebraic relationship between the parameters, satisfying the mathematical structure of the model. Mapping unconstrained continuous embeddings to such structured parameter spaces is not straightforward.

A second approach to the lack of a differentiable parameterization for the posterior of discrete latent variables is to use a continuous relaxation of the discrete distribution. This substitution enables gradient-based optimization, as the continuous formulation allows derivatives to be computed with respect to model parameters. By selecting a relaxation that includes a tunable temperature or smoothness parameter, one can control the extent to which the continuous approximation resembles a discrete distribution, gradually approaching discreteness as the parameter is annealed. More formally, instead of working with a discrete posterior,

$$q_\zeta(\theta \,|\, \mathbf{x}), \qquad \theta \in \{1,\ldots,C\}, \tag{12}$$

we define a relaxed posterior $q_\zeta(\mathbf{z} \mid \mathbf{x})$ over a continuous vector $\mathbf{z} = (z_1,\ldots,z_C)$. This makes the approximate posterior differentiable with respect to the inference model parameters. To motivate the specific choice of relaxation, consider the Gumbel–Max transformation (Gumbel, 1954), which

https://doi.org/10.1017/psy.2026.10105 Published online by Cambridge University Press

expresses a categorical posterior with class probabilities

$$\pi_c(\mathbf{x}) = q_\zeta(\theta = c \mid \mathbf{x}) \tag{13}$$

in terms of a maximization over log probabilities with additive noise:

$$\tilde{\theta} = \underset{c \in \{1,\ldots,C\}}{\operatorname{argmax}} \left(g_c + \log \pi_c(\mathbf{x})\right), \tag{14}$$

where $g_1, \ldots, g_C$ are independent samples from the Gumbel distribution. The fact that the argmax operator is non-differentiable highlights the problem with a categorical posterior for gradient-based estimation. To obtain a differentiable posterior, the argmax operator is replaced with a multinomial logistic function and a temperature parameter λ is introduced. A sample from the relaxed continuous posterior $q_\zeta(\mathbf{z} \mid \mathbf{x})$ is then obtained from

$$\tilde{z}_c = \frac{\exp\left((\log \alpha_c(\mathbf{x}) + g_c)/\lambda\right)}{\sum_{k=1}^{C} \exp\left((\log \alpha_k(\mathbf{x}) + g_k)/\lambda\right)}, \quad c = 1, \ldots, C. \tag{15}$$

Here, $\boldsymbol{\alpha}_\zeta(\mathbf{x}) = (\alpha_1(\mathbf{x}), \ldots, \alpha_C(\mathbf{x}))$ are unnormalized estimates of the posterior class probabilities produced by the inference model, and $\lambda > 0$ controls the smoothness of the relaxation. The resulting vector $\tilde{\mathbf{z}} = (\tilde{z}_1, \ldots, \tilde{z}_C)$ is a relaxed continuous sample from the Concrete distribution: $\mathbf{z} \sim \text{Concrete}(\boldsymbol{\alpha}_\zeta(\mathbf{x}), \lambda)$. As $\lambda \to 0$, the relaxed posterior converges to the original categorical posterior; for larger λ, it becomes smoother and more suitable for gradient-based optimization.

This continuous relaxation of a discrete distribution, which we will refer to as the *concrete distribution*, has been independently proposed by Maddison et al. (2016) who use it for image recognition and density estimation and by Jang et al. (2016) who use it for structured output prediction and semi-supervised learning of generative VAEs. Both authors show that this continuous relaxation of the discrete distribution allows for low-variance unbiased estimates of the gradients with respect to the inference model parameters, while approaching the true discrete nature of the latent variables as the temperature decreases. We can use this same approach to estimate discrete latent variable models by using a concrete distribution as the approximate posterior in our AVI model. We will refer to these models as *concrete latent variable models* (CLVMs).

4. Verifying simulation: LCA and GDINA

Although our key interest is in hybrid models that include both discrete and continuous latent variables, we first demonstrate that the CLVM and ELVM approaches presented above are viable for traditional LCA models. We also consider the GDINA model, as it is a popular discrete latent variable model that can be seen as a constrained LCA model. Specifically, a traditional LCA model can be written as

$$p(\mathbf{x}) = \sum_{c=1}^{C} p(\theta = c) \prod_{i}^{I} P(X_j = x_j | \theta = c), \tag{16}$$

where $\mathbf{x}$ is the vector of binary responses as before, C is the number of latent classes, and I is the number of items. Lazarsfeld (1950) laid the mathematical foundation for the model, but it was first successfully implemented using MML and the EM algorithm by Goodman (1974). We refer to Magidson et al. (2020) for a more comprehensive discussion on the estimation of different variations of the latent class model and the issues that may arise.

The GDINA model (de la Torre, 2011) is a general framework for cognitive diagnostic modeling that includes various popular models. GDINA assumes that K binary latent attributes a_k underly the item scores, and that a known Q-matrix specifies which attributes are required by each item. Specifically, the

https://doi.org/10.1017/psy.2026.10105 Published online by Cambridge University Press

item response function is defined as

$$g[p(X_j = x_j)] = \delta_{j0} + \sum_{k=1}^{K_j} \delta_{jk} a_k q_{jk} + \sum_{k=1}^{K_j - 1} \sum_{k'=k+1}^{K_j} \delta_{jkk'} a_k q_{jk} a_{k'} q_{jk'} + \cdots + \delta_{j12\ldots K_j} \prod_{k=1}^{K_j} a_k q_{jk}, \tag{17}$$

where δ_{j0} is the intercept that sets the guessing probability for the j-th item, δ_{jk} is the coefficient for the effect of attribute k on item j, $\delta_{jkk'}$ is the coefficient for the interaction effect of attributes k and k', and similarly $\delta_{jk_1k_2\ldots k_J}$ is the J-order interaction effect of attributes $a_{k_1}, \ldots, a_{k_J}$. The function g represents a link function that is similar to the link functions from generalized linear modeling, and q_{jk} are elements of the known binary Q indicator matrix. This framework encompasses many different cognitive diagnostic models, and in practice, models are more constrained, determining which order of interaction effects are allowed within each model, and constraining several effects to zero based on expert knowledge. In the current article, we use the identity link function and allow for interactions up to a maximum of two attributes. Additionally, the intercept is constrained to zero. This leaves us with the more concise item response function

$$p(X_j = x_j) = \sum_{k=1}^{K_j} \delta_{jk} a_k + \sum_{k=1}^{K_j - 1} \sum_{k'=k+1}^{K_j} \delta_{jkk'} a_k a_{k'}. \tag{18}$$

4.1. Methods

We simulated data from LCA and GDINA models and estimated parameters using existing MML estimation routines, VI estimation routines and our newly proposed AVI estimation routine. For both models, we simulated datasets from models with either 2 or 10 latent classes/attributes, and with either 20 or 100 items, resulting in four combinations of items and latent dimensions per model. For each combination, we simulated 100 datasets of 10,000 observations each. For LCA, class memberships were sampled from a uniform categorical distribution. In order to have a similar conditional probability structure across classes, we sampled base values for each item from $\text{uniform}(0.3, 0.7)$, and added class-specific noise to these probabilities sampled from $\text{uniform}(-0.2, 0.2)$. In the GDINA model, each attribute is sampled from a Bernoulli distribution with a probability of 0.5. In order to impose structure on the model, we allow each item to load on a maximum of three attributes for the 10 attribute model, and to a maximum of two attributes for the two attribute model. In addition to the main effects, we also include second-order interaction effects of these attributes.[1]

Parameters were then estimated using CLVMs with 1, 5, and 10 importance-weighted samples. For the LCA, the parameters were also estimated using an ELVM with a single sample.[2] This approach was not used for GDINA as the ELVM does not lend itself well for the GDINA model (see Section 3). Both models were also estimated using MML with one and five sets of starting values, respectively. Using multiple sets of starting values is common practice in latent class models as MML for these models is prone to local minima, this practice is reflected in R packages for LCA (Linzer & Lewis, 2011) and GDINA (Ma & de la Torre, 2020) using multiple sets of starting values by default. Finally, we also estimate the models using VI with conjugate prior specification.

For the LCA data, we use both CLVM and ELVM. In the CLVMs, the inference model produces unnoramlized class probabilities $(\alpha_1(\mathbf{x}), \ldots \alpha_C(\mathbf{x}))$, which parameterize the approximate posterior distribution $\text{Concrete}(\boldsymbol{\alpha}(\mathbf{x}), \lambda)$, providing a continuous relaxation of the categorical distribution. In the ELVM approach, the inference model produces a continuous vector $e_\zeta(\mathbf{x})$ which parameterizes the

[1] The exact loading structure used is available on GitHub (Supplementary material).

[2] Using multiple importance-weighted samples for the ELVM would not be meaningful as the approximate posterior for this model is deterministic.

https://doi.org/10.1017/psy.2026.10105 Published online by Cambridge University Press

deterministic approximate posterior of Equation (11). For the standard VI approach, we specify an approximate posterior $q_\lambda(z) = \text{Cat}(\boldsymbol{\pi})$ over class memberships. The parameters $\boldsymbol{\pi}$ are updated directly using the expected log mixture weights and item probabilities, yielding a closed-form variational EM procedure.

For the GDINA model, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ consists of K binary latent attributes. In the CLVM formulation, the inference model outputs unnormalized probabilities

$$\boldsymbol{\phi}_\zeta(\mathbf{x}) = (\phi_1(\mathbf{x}), \ldots, \phi_K(\mathbf{x})), \tag{19}$$

where each $\phi_k(\mathbf{x})$ is a two-dimensional logit vector that parameterizes a concrete distribution (15) for each attribute

$$q_\zeta(\mathbf{z}|\mathbf{x}) = \prod_{k=1}^{K} \text{Concrete}(\boldsymbol{\phi}_k(\mathbf{x}), \lambda), \tag{20}$$

where $\mathbf{z}_k \in (0,1)$ is the continuous relaxation of the binary attribute θ_k. This provides a differentiable relaxation of the factorized Bernoulli distribution which we use in the standard variational GDINA. For comparison, standard VI specifies an approximate posterior $q(\boldsymbol{\theta} \mid \mathbf{x}) = \prod_{k=1}^{K} \text{Bernoulli}(\theta_k \mid \pi_k)$. The parameters $\boldsymbol{\pi}$ are updated directly based on the expected log mixture weights and item response probabilities, which yields a closed-form variational EM procedure, with π updated by conjugacy and the item parameters ψ optimized by gradient ascent.

All AVI- and VI-based models in this article were implemented in PyTorch (Paszke et al., 2019), and the code is publicly available on GitHub (Supplementary material), along with an easy-to-use tool for fitting these models to new datasets.[3] Parameters were estimated using the AMSgrad algorithm for stochastic gradient descent (Reddi et al., 2019), we used a batch size of 1,000 and a learning rate of 0.01, based on our experience, these settings balance stable gradient estimates and sufficiently fast optimization. In general, performance is relatively insensitive to these settings, but setting the batch size to very small values can result in unstable gradient estimates, and choosing a learning rate that is too high might result in suboptimal parameter estimates. We used DReG estimators for the gradients to the inference network. For all models, we started training with a concrete temperature of 1, decreasing it by 10% each iteration. We decrease the temperature to a minimum of 0.01. We chose this minimum threshold because experience showed that decreasing the temperature below 0.01 would result in numerical problems in some cases. We stopped iterating when the IW-ELBO does not decrease by more than 1×10^{-8} for 20 iterations, having these iterations of "patience" rather than immediately stopping when the loss does not decrease is common practice in stochastic optimization of neural networks (Hussein & Shareef, 2024). We set the threshold to an arbitrarily small value for which differences in the ELBO do not reflect meaningful differences in performance. Our choice for the minimum threshold is relatively low compared to other research (e.g., Bouchattaoui et al., 2023; Lin et al., 2025; Oestreich et al., 2024), which means we risk slightly longer training times in order to ensure finding the true minimum. We used a two-layer neural network as the inference model where the number of nodes of the hidden layer is equal to $\frac{I+C}{2}$, the number of layers and the number of nodes were chosen based on earlier research by Urban and Bauer (2021), who use the same values to estimate multidimensional IRT (MIRT) models and note that the performance is relatively insensitive to these values. We also adopt the use of the ELU activation function from the same study.

4.2. Results

Figure 1 shows boxplots containing the parameter recovery results in terms of the mean squared error (MSE) averaged across all conditional probabilities. Furthermore, the right panel of Figure 1 shows the runtime for all methods. Table 2 provides the exact values of these two metrics, along with the quality of latent class estimates as summarized by the average latent class accuracy.

[3] https://github.com/KarelVeldkamp/Discrete_VAEs

https://doi.org/10.1017/psy.2026.10105 Published online by Cambridge University Press

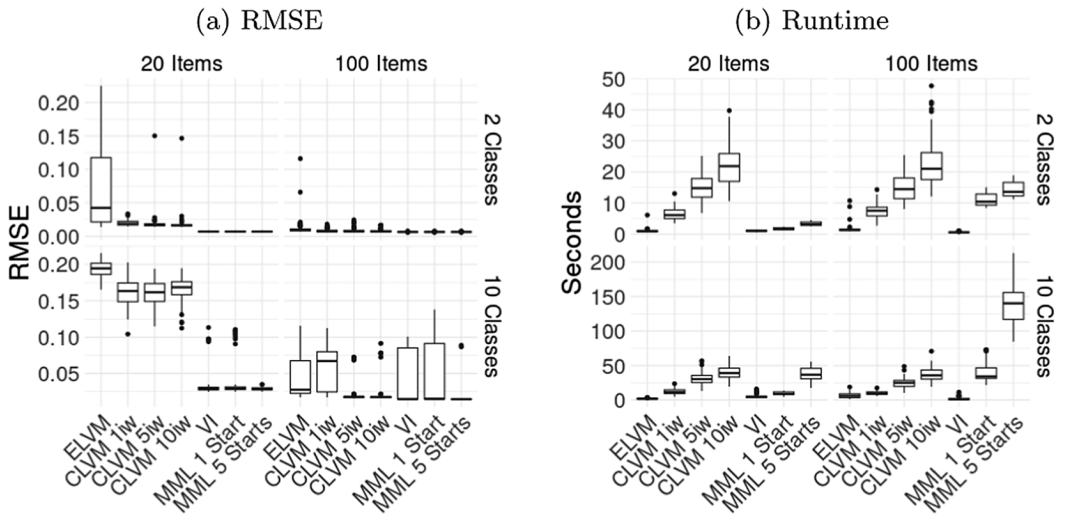

Figure 1. Average RMSE of conditional probabilities (1a) and average runtimes (1b) for LCA across replications, shown for all combinations of the number of items and latent classes. The *x*-axis indicates the estimation method: CLVM with 1, 5, and 10 importance-weighted samples, VI, or MML with 1 or 5 sets of starting values.

Table 2. Root mean squared error (RMSE) of conditional probabilities, accuracy of the latent classes, and runtime in seconds

	20 items			100 items		
Method	RMSE	Accuracy	Runtime	RMSE	Accuracy	Runtime
2 classes						
ELVM	0.082	0.739	1	0.017	0.976	2
CLVM 1iw	0.020	0.873	6	0.009	0.981	8
CLVM 5iw	0.023	0.874	15	0.009	0.980	15
CLVM 10iw	0.022	0.875	22	0.008	0.983	22
VI	0.007	0.882	1	0.007	0.994	1
MML 1 Start	0.007	0.882	2	0.007	0.994	11
MML 5 Starts	0.007	0.882	3	0.007	0.994	14
10 classes						
ELVM	0.195	0.241	2	0.050	0.853	6
CLVM 1iw	0.162	0.300	12	0.065	0.813	10
CLVM 5iw	0.161	0.317	31	0.022	0.925	25
CLVM 10iw	0.167	0.301	40	0.024	0.931	37
VI	0.037	0.556	5	0.050	0.941	2
MML 1 Start	0.042	0.554	9	0.067	0.906	39
MML 5 Starts	0.029	0.560	38	0.019	0.972	138

Note: Metrics are reported for all seven estimation methods for each combination of the number of items and the number of latent classes.

https://doi.org/10.1017/psy.2026.10105 Published online by Cambridge University Press

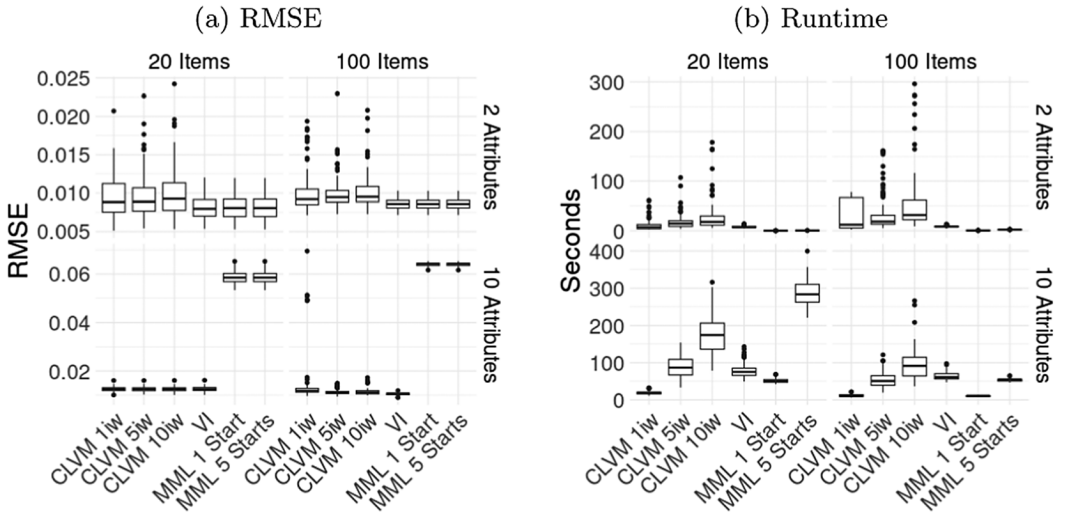

Figure 2. Average RMSE of δ parameters (2a) and average runtimes (2b) for GDINA model across replications, shown for all combinations of the number of items and latent attributes. The *x*-axis indicates the estimation method: CLVM with 1, 5, and 10 importance-weighted samples, VI, or MML with 1 or 5 sets of starting values.

Standard VI is the fastest estimation method for LCA overall, and its parameter estimates are up to par with MML. Although ELVM is the fastest out of the AVI methods, it also provides the worst parameter estimates overall, although it does approach the other methods as the number of items and latent dimensions increase. The RMSE of item parameter estimates is as good as MML with a single start for 100 items and 10 classes, but the latent accuracy is still inferior. CLVM is not as fast as standard VI for LCA, although still faster than MML is some conditions, especially for 10 classes and 100 items, as this condition requires multiple starts in MML. Performance of the CLVM is up to par with MML in most conditions, but it is worse for the condition with 20 items and 10 classes. The effect of using multiple importance weights in the CLVM methods seems limited in the case of LCA, although there is a clear advantage of using multiple importance weights for the condition with 10 classes and 100 items.

For the GDINA model, we report MSE averaged across all item parameters and the average runtime of all methods in Figure 2, and we report the exact values of these metrics along with the average latent attribute accuracy in Table 3. The performance of AVI in terms of accuracy and MSE of the δ parameters is close to MML, except for the conditions with 10 attributes, where MML surprisingly performs worse then both VI and AVI. Interestingly, MML estimation is faster than both VI and AVI methods, even with 10 attributes and 100 items. Standard VI has slightly better latent accuracy than CLVM with a single importance weight.

5. Main simulation: Mixture item response theory

As discussed, our key interest is in models with both discrete and continuous latent variables, for which AVI is expected to have computational advantages over MML. We therefore focus on multidimensional mixture IRT models (Finch & Finch, 2013; Jeon, 2019) which include a discrete latent class variable and multiple continuous latent variables. These models are often used to detect subgroups with different item response strategies or to detect violations of measurement invariance in IRT models (Sen & Cohen, 2019). In our simulation, we focus on a multidimensional IRT model with a mixture on the intercepts. This model assumes that there are two classes $C = 2$, with class probabilities π_c. The item response

https://doi.org/10.1017/psy.2026.10105 Published online by Cambridge University Press

Table 3. Root mean squared error (RMSE) of slopes, accuracy of the latent attributes, and runtime in seconds

Method	RMSE	Accuracy	Runtime	RMSE	Accuracy	Runtime
	20 items			100 items		
2 classes						
CLVM 1iw	0.019	0.997	11	0.021	0.996	28
CLVM 5iw	0.010	1.000	18	0.010	1.000	34
CLVM 10iw	0.011	1.000	28	0.010	1.000	56
VI	0.008	1.000	8	0.009	1.000	9
MML 1 Start	0.008	1.000	<1	0.009	1.000	<1
MML 5 Starts	0.008	1.000	<1	0.009	1.000	2
10 classes						
CLVM 1iw	0.013	1.000	19	0.018	0.996	12
CLVM 5iw	0.013	1.000	89	0.011	1.000	54
CLVM 10iw	0.013	1.000	175	0.011	1.000	96
VI	0.013	1.000	78	0.011	1.000	64
MML 1 Start	0.059	1.000	84	0.064	1.000	9
MML 5 Starts	0.059	1.000	410	0.064	1.000	46

Note: Metrics are reported for all six estimation methods for each combination of the number of items and the number of latent attributes.

function is defined as

$$P(X_j = 1 \mid \boldsymbol{\xi}) = \sum_{c=1}^{C} \pi_c \frac{\exp(\mathbf{a}_j^\top \boldsymbol{\xi} + b_{jc})}{1 + \exp(\mathbf{a}_j^\top \boldsymbol{\xi} + b_{jc})}, \tag{21}$$

where $\boldsymbol{\xi}$ is the vector of continuous latent abilities, $\mathbf{a}_j$ is the vector of slope parameters for item j, b_{jc} is the intercept of item j in class c, and $\pi_c = P(z_c = 1)$ denotes the probability of membership in class c. Estimating separate intercepts for each class allows us to check whether there are differences in item intercepts between unobserved groups of respondents (Sen & Cohen, 2019).

5.1. Methods

We simulate datasets using 3 and 10 continuous latent dimensions for each latent class. In both conditions, we constrain the item slopes in such a way that each item loads on only one or two latent dimensions. We use 28 items for the three-dimensional model and 110 items for the 10-dimensional model. The exact loading structure is available on GitHub (Supplementary material). We simulate 100 datasets for each condition, each consisting of 10,000 observations, matching the dataset size of our real data application. The class memberships are drawn with a probability of 0.5, and the item slopes were drawn from $U(0.5, 2)$. The overall intercepts are drawn from $U(-0.2, 0.2)$, we then add standard normal noise to the intercepts of one of the classes to ensure that the intercepts vary across the two classes. For each dataset, we estimate the model using a CLVM with either 1, 5, or 10 importance-weighted samples, and we estimate the model using MML as implemented in Mplus (Muthén & Muthén, 2017) using either a single set of starting values or using five sets of starting values.

https://doi.org/10.1017/psy.2026.10105 Published online by Cambridge University Press

For the CLVM, the approximate posterior is defined over both the discrete class variable and the continuous latent abilities:

$$q_\zeta(\boldsymbol{\theta} \mid \mathbf{x}) = q_\zeta(\mathbf{z} \mid \mathbf{x})\, q_\zeta(\boldsymbol{\xi} \mid \mathbf{x}, \mathbf{z}), \tag{22}$$

where $\mathbf{z}$ denotes the discrete latent class indicator and $\boldsymbol{\xi}$ denotes the continuous latent abilities. Thus, in the mixture IRT setting, $\boldsymbol{\theta}$ comprises both discrete and continuous components $\boldsymbol{\theta} = (\boldsymbol{\zeta}, \boldsymbol{\xi})$. Similar to LCA and GDINA, the inference network outputs unnormalized class logits which parameterize a Concrete distribution (15) In the standard VI setting, this distribution is replaced by a categorical distribution without relaxation. Conditioned on the class assignment, the continuous latent abilities are modeled using a class-conditional diagonal Gaussian. The inference network also outputs class-specific mean and log-variance parameters:

$$\boldsymbol{\mu}_\zeta(\mathbf{x}) = (\boldsymbol{\mu}_1(\mathbf{x}), \ldots, \boldsymbol{\mu}_K(\mathbf{x})), \qquad \log \boldsymbol{\sigma}^2_\zeta(\mathbf{x}) = (\log \boldsymbol{\sigma}^2_1(\mathbf{x}), \ldots, \log \boldsymbol{\sigma}^2_K(\mathbf{x})), \tag{23}$$

where each pair $(\boldsymbol{\mu}_k(\mathbf{x}), \log \boldsymbol{\sigma}^2_k(\mathbf{x}))$ parameterizes a Gaussian component

$$q_\zeta(\boldsymbol{\xi} \mid \mathbf{x}, \mathbf{z} = k) = \mathcal{N}\big(\boldsymbol{\mu}_k(\mathbf{x}), \operatorname{diag}(\boldsymbol{\sigma}^2_k(\mathbf{x}))\big). \tag{24}$$

Under the Concrete relaxation, where $\mathbf{z}$ is a relaxed one-hot vector, the Gaussian parameters are computed as

$$\boldsymbol{\mu}_\zeta(\mathbf{x}, \mathbf{z}) = \sum_{k=1}^{K} z_k \boldsymbol{\mu}_k(\mathbf{x}), \qquad \log \boldsymbol{\sigma}^2_\zeta(\mathbf{x}, \mathbf{z}) = \sum_{k=1}^{K} z_k \log \boldsymbol{\sigma}^2_k(\mathbf{x}) \tag{25}$$

so that

$$q_\zeta(\boldsymbol{\xi} \mid \mathbf{x}, \mathbf{z}) = \mathcal{N}\big(\boldsymbol{\mu}_\zeta(\mathbf{x}, \mathbf{z}), \operatorname{diag}(\boldsymbol{\sigma}^2_\zeta(\mathbf{x}, \mathbf{z}))\big). \tag{26}$$

5.2. Results

Figure 3 presents the MSE averaged over the intercept parameters for the different estimation methods in the left panel, as well as the average runtimes in the right panel. Table 4 contains the exact MSE for both the slopes and intercepts, as well as the average latent class accuracy and the average runtimes. Note that we only provide MML estimation results for the three-dimensional model, as it is numerically infeasible to estimate a 10-dimensional mixture model using any reasonable number of quadrature points. The results show that both for VI and AVI, using multiple importance-weighted samples generally improves the parameter estimates, although using five importance weights rather than 10 sometimes seems to result in slightly better results for the three-dimensional model. The MSE of the parameter estimates of the AVI- and VI-based methods with multiple importance weights is slightly better than the MML estimates, whereas MML results in slightly better latent class accuracy. For the mixture IRT models, AVI and VI methods are both clearly much faster than MML, with AVI being even faster than VI. For the three-dimensional model, AVI with five importance-weighted samples was about 15 times faster than MML with a single start, and nine times faster than standard VI with the same number of importance weights. For the 10-dimensional model, AVI was over twice as fast as standard VI, while we were unable to obtain good estimates using MML in any reasonable amount of time.

6. Application: Narcissism personality inventory

To illustrate the utility of our approach in a real data application, we apply a mixture-IRT model to a dataset of responses to the Narcissistic Personality Inventory, which is a questionnaire measuring narcissism developed by Raskin and Terry (1988). This inventory consists of 40 binary forced-choice questions relating to narcissism. The inventory is split into seven subscales, namely, Authority, Entitlement, Exhibitionism, Exploitativeness, Superiority, Self-Sufficiency, and Vanity. The dataset we use is

https://doi.org/10.1017/psy.2026.10105 Published online by Cambridge University Press

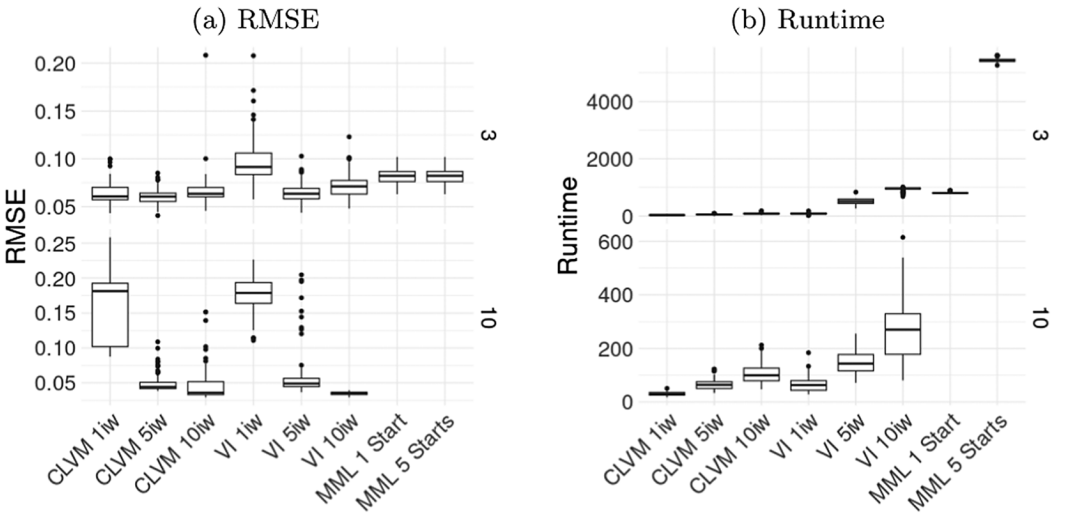

Figure 3. Average RMSE of the intercepts (3a) and average runtimes in seconds (3b) across replications for 3- and 10-dimensional models, respectively. The *x*-axis indicates the estimation method: CLVM with 1, 5, and 10 importance-weighted samples, VI or MML with 1 or 5 sets of starting values.

Table 4. RMSE of intercepts (b) and slopes (a), accuracy of the latent classes, and runtime in seconds for the mixture IRT simulation

Method	RMSE(b)	RMSE(a)	Accuracy	Runtime
3 dimensions				
CLVM 1iw	0.065	0.133	0.947	36
CLVM 5iw	0.061	0.099	0.951	56
CLVM 10iw	0.068	0.140	0.950	87
VI 1iw	0.099	0.335	0.949	88
VI 5iw	0.065	0.101	0.956	506
VI 10iw	0.074	0.111	0.956	939
MML 1 Start	0.082	0.182	0.958	812
MML 5 Starts	0.082	0.182	0.958	5,440
10 dimensions				
CLVM 1iw	0.170	0.116	0.985	31
CLVM 5iw	0.051	0.102	0.987	65
CLVM 10iw	0.049	0.113	0.989	104
VI 1iw	0.177	0.514	0.884	67
VI 5iw	0.069	0.340	0.969	146
VI 10iw	0.035	0.098	0.994	265

Note: Metrics are reported for all estimation methods for both the 3- and 10-dimensional models. Note that MML was only feasible for the three-dimensional model.

https://doi.org/10.1017/psy.2026.10105 Published online by Cambridge University Press

collected through a convenience sample obtained from the open-source psychometrics project,[4] and consists of 11,243 respondents (6,425 Male, 4,766 Female, 40 other, 12 Missing). The average age of participants was 34.01 (SD 15.02).

In order to model the heterogeneity in narcissism across participants, we estimate a seven-dimensional mixture-IRT model. Note that estimating such a model using quadrature-based MML would be infeasible. Existing latent variable estimation software recommends using 10–20 quadrature points per dimension (Muthén & Muthén, 2017; Vermunt & Magidson, 2021), which would result in over 10 million quadrature points. We use a simple structure for the factor loadings, which corresponds to the seven components identified by Raskin and Terry (1988). The exact latent structure is available on GitHub (Supplementary material). We constrain the slope parameters to be equal across classes, whereas the intercepts are class-specific. We code the items so that a one denotes that the narcissistic option was selected and a zero denotes that the non-narcissistic option was selected. The coding for each item is available in the codebook included with the open-psychometrics dataset. By allowing the intercepts to vary across two latent classes, we can investigate differences in propensity to narcissism across discrete latent classes. We fit two models, one with a single latent class and one with two latent classes, and determine which model fits the data best using the cross-validated log-likelihood. The AVI approach allows us to estimate such a model in less than 20 seconds, which also allows us to compute bootstrapped standard errors for the parameter estimates.

We estimate the models using a CLVM with 10 importance-weighted samples, which takes less than 20 seconds. The cross-validated log-likelihood for the two-class model was higher (−35,506.562) than for the model with a single class (−36,589.54), indicating that the two-class model is better suited for the data at hand, reflecting the heterogeneity in the data. All other results in this section are based on the two-class model.

The average non-zero slope estimate in the two-class model was 1.22 (SD=0.68), and the slope estimates fall within a reasonable range, as reflected in the boxplot in Figure 4a. More importantly, Figure 4b plots the intercepts of both classes against each other. Although the intercepts of the two classes are strongly correlated ($\rho = 0.843$), the intercepts in class 2 are clearly higher overall. Our two-class model does not suggest certain item groups that function differently across latent groups, but rather an overall difference between classes in the propensity to answer in a narcissistic fashion. Figure 4c shows the 95% bootstrapped confidence intervals for the difference in intercept between classes for each item. This confirms that subjects in class two are more likely to answer in a narcissistic fashion than people in class one.

Among males, the probability of being in the more narcissistic class is 0.47, whereas the probability of females being in the narcissistic class is only 0.34. This in line with the literature on narcissism, which consistently finds that males are more likely to possess this trait (Grijalva et al., 2015). The average age in the narcissistic class is 30.42 (SD=15.51) and the average age in the other class is 36.54 (SD=16.08), indicating that younger subjects are more likely to be in the narcissistic class, which is also consistent with the literature, as narcissism is known to be most prevalent in early adulthood and decrease linearly with age (Weidmann et al., 2023).

7. Discussion

We studied the performance of AVI in estimating models with discrete latent variables. In an initial simulation, we verified that AVI can be used to accurately estimate parameters for traditional discrete latent variable models, such as LCA and GDINA. However, although AVI was faster than MML in some conditions for LCA, there is no clear computational advantage for LCA and GDINA models overall, and standard VI is actually computationally more efficient in these cases due to the closed-form analytical updates of the approximate posterior. In our main simulation, we applied AVI to a more

[4] https://openpsychometrics.org/_rawdata/

https://doi.org/10.1017/psy.2026.10105 Published online by Cambridge University Press

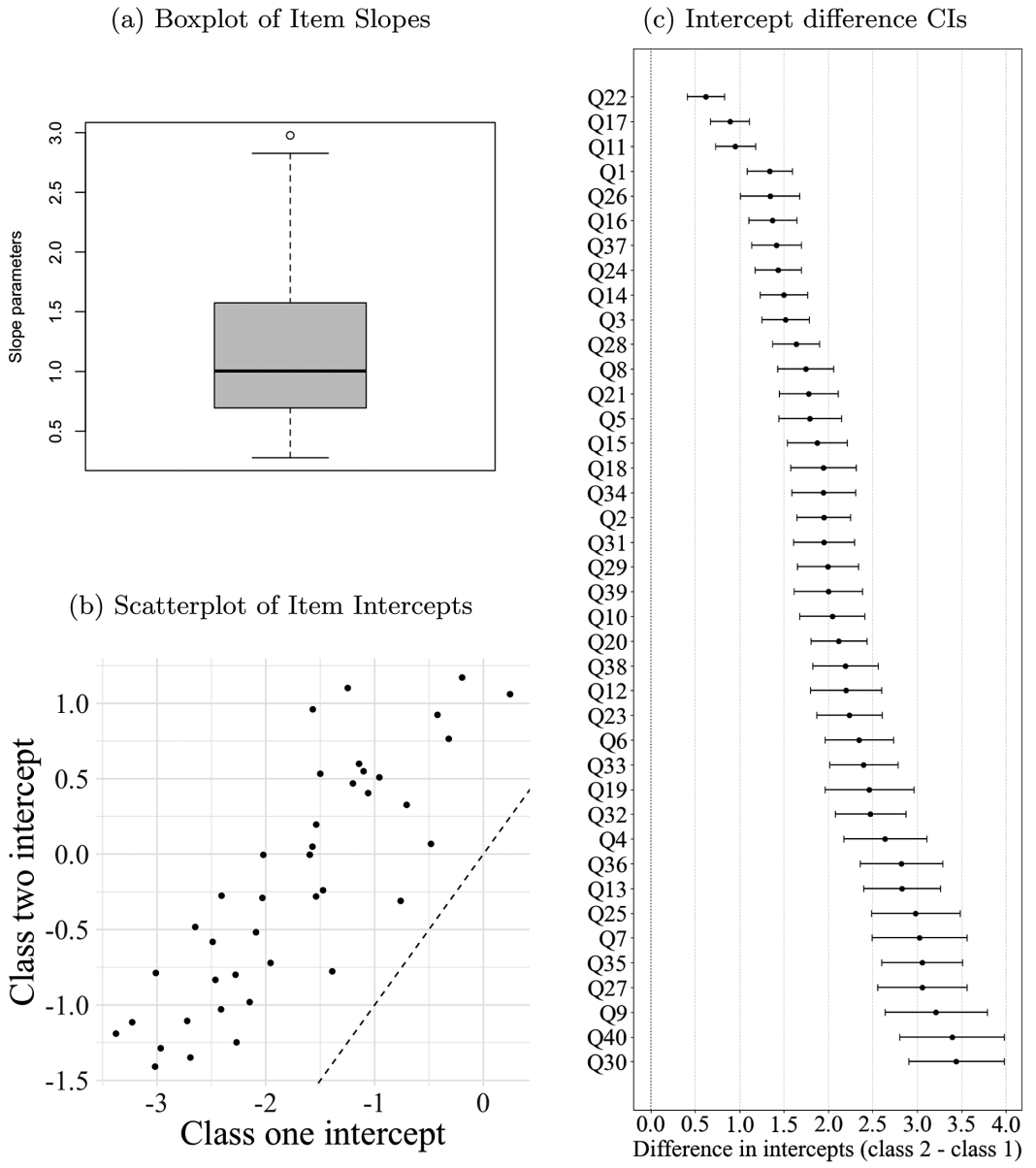

Figure 4. Distribution of slope parameters (4a), scatterplot of class intercepts (4b), and 95% confidence intervals of the difference in intercepts across classes (4c).

complex mixture multidimensional IRT model. Results indicate that AVI with a concrete distribution as the approximate posterior distribution results in accurate parameter estimates for mixture IRT models, while being much faster than MML and standard VI estimation. Our AVI approach even allowed to estimate multi-dimensional mixture IRT models that were too complex for MML estimation. Furthermore, we demonstrated the practical applicability of our approach in an application to the NPI, estimating a seven-dimensional mixture IRT model very efficiently, even allowing for bootstrapped standard error estimates.

https://doi.org/10.1017/psy.2026.10105 Published online by Cambridge University Press

The finding that AVI was computationally slower than MML for GDINA models was unexpected. A possible explanation for this result is the way that the MML optimization routine exploits the structure of the GDINA model much better than our AVI approach. The MML estimate of each item parameter can be expressed in terms of the expected count of correct responses and the number of individuals with an attribute pattern (see de la Torre, 2011, Equation (15)). Because the Q-matrix specifies which items load on each attribute in GDINA, these updates can be performed independently for each item, so the estimation proceeds in low-dimensional item-specific updates that do not require numerical gradient optimization. In the AVI approach, all model parameters are updated jointly, which requires sampling from the approximate posterior and computing gradients for all item parameters.

Although this research demonstrates the use of AVI for discrete latent variables generally, the application currently remains limited to LCA, GDINA, and mixture IRT. However, there are more discrete latent variable models that could potentially benefit from this approach, such as factor mixture models (Kim et al., 2023) or hidden Markov models (Mor et al., 2021). Future research is needed to extend the AVI approach to a broader set of models and to test its computational advantage in other fields.

Another avenue for research concerns the setting of different hyperparameters, such as the number of layers in the inference model, the number of nodes in each layer, the activation functions used, the learning rate, and the temperature decay. We use settings based on previous research (Urban & Bauer, 2021) and our own experience, which seem to work well in a wide range of settings. However, a more systematic investigation of the effect of these hyperparameters in different settings and for different models would be valuable, although this is beyond the scope of the current research. Generally, in practice, one can adopt the settings used in this study, but if the data or models are highly different from the ones considered here, we recommend using cross-validation to specify these settings.

Overall, our research shows that AVI can successfully be generalized to discrete latent variable models. The approach can provide accurate parameter estimates in a broad set of discrete latent variable models, but is most beneficial for complex latent variable models where there are both continuous and discrete latent variable models, as the computational advantages are most pronounced for these models.

Funding statement. Open access funding provided by University of Amsterdam

Supplementary material. The supplementary material for this article can be found at https://github.com/KarelVeldkamp/Discrete_VAEs.

Acknowledgments. All data, parameters, code and complete results are available in the supplementary material. AI tools were used for minor writing suggestions. The authors take full responsibility for the content of this article.

Funding statement. This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Competing interests. The authors declare no competing interests.

References

Bartholomew, D. J., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach*. John Wiley & Sons.

Bengio, Y., Léonard, N., & Courville, A. (2013). *Estimating or propagating gradients through stochastic neurons for conditional computation*. Preprint, arXiv:1308.3432.

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

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

Bouchattaoui, M. E., Tami, M., Lepetit, B., & Cournède, P.-H. (2023). *Causal dynamic variational autoencoder for counterfactual regression in longitudinal data*. Preprint, arXiv:2310.10559.

Burda, Y., Grosse, R., & Salakhutdinov, R. (2015). *Importance weighted autoencoders*. Preprint, arXiv:1509.00519.

Cai, L. (2010). High-dimensional exploratory item factor analysis by a metropolis–Hastings Robbins–Monro algorithm. *Psychometrika*, *75*, 33–57.

https://doi.org/10.1017/psy.2026.10105 Published online by Cambridge University Press

Chen, Y., Li, X., & Zhang, S. (2019). Joint maximum likelihood estimation for high-dimensional exploratory item factor analysis. *Psychometrika*, *84*, 124–146.

Cho, A. E., Wang, C., Zhang, X., & Xu, G. (2021). Gaussian variational estimation for multidimensional item response theory. *British Journal of Mathematical and Statistical Psychology*, *74*, 52–85.

Converse, G., Curi, M., Oliveira, S., & Templin, J. (2021). Estimation of multidimensional item response theory models with correlated latent variables using variational autoencoders. *Machine Learning*, *110*(6), 1463–1480.

Curi, M., Converse, G. A., Hajewski, J., & Oliveira, S. (2019). Interpretable variational autoencoders for cognitive models. In Chrisina Jayne; Zoltán Somogyvári (Eds.). *2019 international joint conference on neural networks (IJCNN)* (pp. 1–8), IEEE.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*(2), 179–199.

de la Torre, J., & Minchen, N. (2014). Cognitively diagnostic assessments and the cognitive diagnosis model framework. *Psicolog í a Educativa*, *20*(2), 89–97.

Debelak, R., & Urban, C. J. (2021). *An evaluation of deep learning approaches for factor analysis of response and response time data.* Preprint.

Edwards, M. C. (2010). A Markov chain Monte Carlo approach to confirmatory item factor analysis. *Psychometrika*, *75*(3), 474–497.

Finch, W. H., & Finch, M. E. H. (2013). Investigation of specific learning disability and testing accommodations based differential item functioning using a multilevel multidimensional mixture item response theory model. *Educational and Psychological Measurement*, *73*(6), 973–993.

Formann, A. K., & Kohlmann, T. (1996). Latent class analysis in medical research. *Statistical Methods in Medical Research*, *5*(2), 179–211.

Goodman, L. A. (1974). The analysis of systems of qualitative variables when some of the variables are unobservable. Part I—A modified latent structure approach. *American Journal of Sociology*, *79*(5), 1179–1259.

Grijalva, E., Newman, D. A., Tay, L., Donnellan, M. B., Harms, P. D., Robins, R. W., & Yan, T. (2015). Gender differences in narcissism: A meta-analytic review. *Psychological Bulletin*, *141*(2), 261.

Grimmer, J. (2011). An introduction to Bayesian inference via variational approximations. *Political Analysis*, *19*(1), 32–47.

Gumbel, E. J. (1954). *Statistical theory of extreme values and some practical applications*, National Bureau of Standards Applied Mathematics Series, 33. U.S. Government Printing Office.

Hasan, M., Deng, L. Y., Sabatini, J., Bowman, D., Yang, C.-C., & Hollander, J. (2022). Effect of Q-matrix misspecification on variational autoencoders (VAE) for multidimensional item response theory (MIRT) models estimation. In Antonija Mitrovic; Nigel Bosch (Eds). *Proceedings of the 15th international conference on educational data mining* (p. 811). International Educational Data Mining Society.

Hussein, B. M., & Shareef, S. M. (2024). An empirical study on the correlation between early stopping patience and epochs in deep learning. In *ITM web of conferences* (Vol. *64*, p. 1003). EDP Sciences.

Jang, E., Gu, S., & Poole, B. (2016). *Categorical reparameterization with Gumbel-Softmax*. Preprint, arXiv:1611.01144.

Jeon, M. (2019). A specialized confirmatory mixture IRT modeling approach for multidimensional tests. *Psychological Test and Assessment Modeling*, *61*(1), 91–123.

Kahn, H., & Marshall, A. W. (1953). Methods of reducing sample size in Monte Carlo computations. *Journal of the Operations Research Society of America*, *1*(5), 263–278.

Kim, E., Wang, Y., & Hsu, H.-Y. (2023). A systematic review of and reflection on the applications of factor mixture modeling. *Psychological Methods*, *30*(5), 997.

Kingma, D. P., & Welling, M. (2013). *Auto-encoding variational Bayes*. Preprint, arXiv:1312.6114.

Lazarsfeld, P. F. (1950). *The logical and mathematical foundation of latent structure analysis*, Studies in Social Psychology in World War II Vol. IV: Measurement and Prediction (pp. 362–412). Princeton University Press.

Lin, L., Peiro-Corbacho, P., Ávila, P., Carta-Bergaz, A., Arenal, Á., Ríos-Muñoz, G. R., & Sevilla-Salcedo, C. (2025). *CLARAE: Clarity preserving reconstruction autoencoder for denoising and rhythm classification of intracardiac electrograms.* Preprint, arXiv:2510.17821. https://doi.org/10.48550/arXiv.2510.17821

Linzer, D. A., & Lewis, J. B. (2011). poLCA: An R package for polytomous variable latent class analysis. *Journal of Statistical Software*, *42*, 1–29.

Liu, T., Wang, C., & Xu, G. (2022). Estimating three-and four-parameter MIRT models with importance-weighted sampling enhanced variational auto-encoder. *Frontiers in Psychology*, *13*, 4189.

Ma, C., Ouyang, J., Wang, C., & Xu, G. (2023). A note on improving variational estimation for multidimensional item response theory. *Psychometrika*, *89*(1), 172–204.

Ma, W., & de la Torre, J. (2020). GDINA: An R package for cognitive diagnosis modeling. *Journal of Statistical Software*, *93*, 1–26.

Maddison, C. J., Mnih, A., & Teh, Y. W. (2016). *The concrete distribution: A continuous relaxation of discrete random variables*. Preprint, arXiv:1611.00712.

Magidson, J., Vermunt, J. K., & Madura, J. P. (2020). *Latent class analysis*. SAGE Publications Limited.

Molenaar, D., Grasman, R. P., & Cúri, M. (2025). Autoencoders for amortized joint maximum likelihood estimation of confirmatory item factor models. *Multivariate Behavioral Research*, *60*(4), 657–677.

https://doi.org/10.1017/psy.2026.10105 Published online by Cambridge University Press

Mor, B., Garhwal, S., & Kumar, A. (2021). A systematic review of hidden Markov models and their applications. *Archives of Computational Methods in Engineering*, *28*, 1429–1448.

Muthén, L. K., & Muthén, B. O. (2017). Mplus user's guide (8th ed.) [computer software manual]. Version 8.

Oestreich, M., Ewert, I., & Becker, M. (2024). Small molecule autoencoders: Architecture engineering to optimize latent space utility and sustainability. *Journal of Cheminformatics*, *16*(1), 26.

Oka, M., & Okada, K. (2023). Scalable Bayesian approach for the Dina Q-matrix estimation combining stochastic optimization and variational inference. *Psychometrika*, *88*(1), 302–331.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., ... Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.) *Advances in neural information processing systems* (Vol. *32*, pp. 8024–8035). Curran Associates, Inc. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

Radev, S. T., Mertens, U. K., Voss, A., Ardizzone, L., & Köthe, U. (2020). Bayesflow: Learning complex stochastic models with invertible neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(4), 1452–1466.

Rainforth, T., Kosiorek, A., Le, T. A., Maddison, C., Igl, M., Wood, F., & Teh, Y. W. (2018). Tighter variational bounds are not necessarily better. In J. Dy & A. Krause (Eds.), *International conference on machine learning* (pp. 4277–4285).

Raskin, R., & Terry, H. (1988). A principal-components analysis of the narcissistic personality inventory and further evidence of its construct validity. *Journal of Personality and Social Psychology*, *54*(5), 890. PMLR.

Reddi, S. J., Kale, S., & Kumar, S. (2019). *On the convergence of Adam and beyond*. Preprint, arXiv:1904.09237.

Sen, S., & Cohen, A. S. (2019). Applications of mixture IRT models: A literature review. *Measurement: Interdisciplinary Research and Perspectives*, *17*(4), 177–191.

Sullivan, P. F., Kessler, R. C., & Kendler, K. S. (1998). Latent class analysis of lifetime depressive symptoms in the national comorbidity survey. *American Journal of Psychiatry*, *155*(10), 1398–1406.

Tucker, G., Lawson, D., Gu, S., & Maddison, C. J. (2018). *Doubly reparameterized gradient estimators for Monte Carlo objectives*. Preprint, arXiv:1810.04152.

Urban, C. J., & Bauer, D. J. (2021). A deep learning algorithm for high-dimensional exploratory item factor analysis. *Psychometrika*, *86*(1), 1–29.

van den Oord, A., Vinyals, O., & Kavukcuoglu, K. (2017). Neural discrete representation learning. In I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. *30*, pp. 6306–6315). Curran Associates, Inc.

Veldkamp, K., Grasman, R., & Molenaar, D. (2024). Handling missing data in variational autoencoder based item response theory. *British Journal of Mathematical and Statistical Psychology*, *78*(1), 378–397.

Vermunt, J. K., & Magidson, J. (2021). Latent GOLD 6.0 user's guide [computer software manual]. https://www.statisticalinnovations.com

von Davier, M., & Sinharay, S. (2010). Stochastic approximation methods for latent regression item response models. *Journal of Educational and Behavioral Statistics*, *35*(2), 174–193.

Weidmann, R., Chopik, W. J., Ackerman, A. M., RobertA., Bianchi, E. C., Brecheen, C., Campbell, W. K., Gerlach, T. M., Geukes, K., Grijalva, E., Grossmann, I., Hopwood, C. J., Hutteman, R., Konrath, S., Küfner, A. C. P., Leckelt, M., Miller, J. D., Penke, L., Pincus, A. L., ... Back, M. D. (2023). Age and gender differences in narcissism: A comprehensive study across eight measures and over 250,000 participants. *Journal of Personality and Social Psychology*, *124*(6), 1277.

Wu, M., Davis, R. L., Domingue, B. W., Piech, C., & Goodman, N. (2021). *Modeling item response theory with stochastic variational inference*. Preprint, arXiv:2108.11579.

Yamaguchi, K. (2000). Multinomial logit latent-class regression models: An analysis of the predictors of gender-role attitudes among Japanese women. *American Journal of Sociology*, *105*(6), 1702–1740.

Yamaguchi, K., & Okada, K. (2020). Variational Bayes inference algorithm for the saturated diagnostic classification model. *Psychometrika*, *85*(4), 973–995.

Zhang, L., & Chen, P. (2024). A neural network paradigm for modeling psychometric data and estimating IRT model parameters: Cross estimation network. *Behavior Research Methods*, *56*(7), 7026–7058.

Cite this article: Veldkamp, K., Grasman, R. and Molenaar, D. (2026). Estimating Discrete Latent Variable Models Using Amortized Variational Inference. *Psychometrika*, 1–19. https://doi.org/10.1017/psy.2026.10105